Non-Spurious Correlations between Genetic and Linguistic Diversities in the Context of Human Evolution

Dan Dediu

Bsc. Math. Comp. Sci., Msc. Neurobiol. Behav.

A thesis submitted in fulfillment of requirements for the degree of

Doctor of Philosophy

to

Linguistics and English Language

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

February 2007

ALL MISSING PAGES ARE BLANK

IN

ORIGINAL

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgment is made in the text.

Dan Dediu

Abstract

This thesis concerns human diversity, arguing that it represents not just some form of noise, which must be filtered out in order to reach a deeper explanatory level, but the engine of human and language evolution, metaphorically put, the best gift Nature has made to us. This diversity must be understood in the context of (and must shape) human evolution, of which the Recent Out-of-Africa with Replacement model (ROA) is currently regarded, especially outside palaeoanthropology, as a true theory. It is argued, using data from palaeoanthropology, human population genetics, ancient DNA studies and primatology, that this model must be, at least, amended, and most probably, rejected, and its alternatives must be based on the concept of reticulation.

The relationships between the genetic and linguistic diversities is complex, including inter-individual genetic and behavioural differences (behaviour genetics) and inter-population differences due to common demographic, geographic and historic factors (spurious correlations), used to study (pre)historical processes. It is proposed that there also exist non-spurious correlations between genetic and linguistic diversities, due to genetic variants which can bias the process of language change, so that the probabilities of alternative linguistic states are altered. The particular hypothesis (formulated with Prof. D. R. Ladd) of a causal relationship between two human genes and one linguistic typological feature is supported by the statistical analysis of a vast database of 983 genetic variants and 26 linguistic features in 49 Old World populations, controlling for geography and known linguistic history.

The general theory of non-spurious correlations between genetic and linguistic diversities is developed and its consequences and predictions analyzed. It will very probably profoundly impact our understanding of human diversity and will offer a firm footing for theories of language evolution and change. More specifically, through such a mechanism, gradual, accretionary models of language evolution are a natural consequence of post-ROA human evolutionary models.

The unravellings of causal effects of inter-population genetic differences on linguistic states, mediated by complex processes of cultural evolution (biased iterated learning), will represent a major advance in our understanding of the relationship between cultural and genetic diversities, and will allow a better appreciation of this most fundamental and supremely valuable characteristic of humanity – its intrinsic diversity.

Acknowledgments

Many thanks to my supervisors, Jim Hurford and Simon Kirby, for encouragement and many critical and insightful discussions, even when I was attacking controversial issues (which I almost always did). To D. Robert Ladd for sharing his vast knowledge of linguistics and willingness to work hard on the "crazy" idea of non-spurious correlations, and for his continuous support.

To Mónica Tamariz, Stefan Höfler and Dave Hawkey for discussions on linguistic diversity, its meanings and origins, to Anna Parker for help with getting this thesis right and all the other LEC members for sharing ideas and thought-provoking discussions. To Carmen Strungaru and Wulf Schiefenhövel for making me discover language evolution.

Special thanks to my wife, Alexandra Dima, for her loving and funny way of being, and for her invaluable help with statistics, and her decisive role in my actually learning it. I would not have been here without her... To my parents, who made me like so many aspects of Nature, from biology to mathematics, teaching me how to think freely (especially in a time and place when this was forbidden), and helping me pursue my dreams, despite many hardships.

The Overseas Research Students Awards Scheme (ORS Award 2003014001) and the College of Humanities and Social Science, University of Edinburgh's Studentship made this financially possible.

I thank B. Connell, C. Kutsch Lojenga, H. Eaton, J. A. Edmondson, J. Hurford, K. Bostoen, L. Ziwo, M. Blackings, N. Fabb, O. Stegen, R. Asher, R. Ridouane, M. Endl and J. Roberts for help with language data and R. McMahon for comments on an early draft of the tone-genes correlation idea.

Table of Contents

Index of Tables

Index of Figures

1. Introduction and overview

There is sometimes a tendency, when thinking about the evolution of language, to abstract away from the details of human evolution, by making sketchy and unanalyzed assumptions. The Recent Out-of-Africa with Replacement model of human evolution seems to be a *de facto* standard, considered to be true, especially outside palaeoanthropology, but I will argue in Chapter 2 that it has a series of problems, some of them important enough to lead to its falsification. The alternatives are based on the concept of *reticulation*, involving a single (or a very limited number of) species of the genus *Homo*, during its entire evolutionary history and geographical range.

Language evolution must have involved a certain level of biological evolution (not necessarily specific for language), thus, of genetic changes. The nature of the correlations between genes and languages is analyzed in Chapter 3. *Inter-individual differences* in language behaviour are correlated with genetic differences, allowing behavioural genetic approaches (heritability, group heritability, etc.), capable of illuminating the genetic bases of the capacity for language. The most probable model seems to be of many genes with small effects, even though genes with catastrophic effects (like *FOXP2*) are interesting to study. This impacts on the probability of catastrophic macromutations in language evolution, favoring gradual, accretionary models. *Inter-population genetic differences* could also correlate with linguistic differences, in the sense that there are factors (geographic, demographic, historic) which shape both diversities in similar ways. Therefore, when present, these *spurious correlations* can shed light on (pre)historic events, but the methods used and results obtained so far point to the immaturity of this field.

The current assumption of the uniformity of the capacity for language has made impossible inquiries into the existence of *non-spurious correlations between genetic and linguistic diversities*, whereby genetic variants could bias the trajectory of language change, so that they cause changes in the probabilities of certain linguistic states. The discussion of this general theory of non-spurious correlations between genetic and linguistic diversities and its main consequences is contained in Chapter 5, while Chapter 4 analyses a particular case. This case is represented by the hypothesized (together with Prof. D. R. Ladd) relationship between two genes involved in brain growth and development, showing signs of natural

selection and geographic patterning in human populations, *ASPM* and *Microcephalin*, and the typological linguistic feature of *tone*. This hypothesis is tested using a vast database of 983 genetic variants and 26 linguistic features in 49 populations of the Old World, also controlling for geography and known history. The result is that the correlation between them is both statistically significant (0.05 level) and in the top 5% of the empirical distribution of the database (it is much stronger than the vast majority of other such correlations), suggesting that the relationship is real and not due to geographical or historical factors. The methodology developed here, drawing on geo- and spatial statistics, evolutionary biology, population genetics, linguistic typology and classical statistics can be used to answer many questions concerning genetic and linguistic diversities and their relationships, and warrants further refinement.

The theory of non-spurious correlations between genetic and linguistic diversities, if confirmed by future, more demanding and targeted studies, would prove to be a paradigm change not only for linguistics and genetics, but also for human evolution and our general understanding of human diversity. Concerning language evolution, it offers the basis for gradual, accretionary models, whereby genetic and linguistic diversities represent the engines of human and language evolution, and not just some noise which must be dealt with. Chapter 5 also offers a novel framework for language evolution, firmly grounded in human evolution, where such non-spurious correlations are the main explanatory device.

It is hoped that this theory of non-spurious correlations between genetic and linguistic diversities will be confirmed and refined by further study and will offer a new, firmer basis for understanding human diversity, the most important gift Nature has made to us.

2. Human evolution

This chapter will review the current human evolutionary models and the controversies surrounding them. The Recent-Out-of-Africa with Replacement model will be presented, as it is currently perceived to be almost unanimously accepted, focusing on its implications and history. The first usage of mtDNA (in 1987) to answer questions bearing on human origins is also presented, as it offers the opportunity to clarify different evolutionary genetic concepts and methods. The three historical models of human evolution, polygenism, monogenism and multiregionalism, are discussed in their own contexts, trying to dispel the myths surrounding each of them. The chapter then focuses on a series of issues concerning the Recent-Out-of-Africa with Replacement model and concludes with two recent, better alternatives.

As is very well known and generally accepted, the evolution of modern humans is explained by the *Recent Out-of-Africa model* (Stringer & Andrews, 1988; denoted in the following as ROA), whereby we, the sole surviving species of a bushy and specious genus, evolved from some ancestral *Homo* stock in eastern Africa somewhere around 200-150kya[1] and further dispersed throughout the Old World, replacing the local archaic human forms, and, much later, into the New World. Thus, we are a young, very uniform, very versatile and invasive species. Given these, the students of language evolution *seem to have to* assume a uniform biological capacity for language (due to our low diversity) and a recent origins of modern speech (simultaneous with the modern *Homo sapiens* speciation event).

But... is it *true*? (Wildavsky, 1997) Is ROA so firmly established as a valid explanation of our evolutionary emergence? And, if not, why does it appear to be indisputably true, especially outside the palaeoanthropological community? Why is the debate more or less hidden to the outsiders? And, more importantly for us, what are the proposed alternatives? Are they better suited to explain the seemingly extremely complex and messy data we have collected so far? Does this possible shift also offer new approaches and challenges for the students of language evolution?

1 kya = thousands of years ago (i.e., BP – before present).

2.1. The Recent Out-of-Africa Model and the Evolutionary History of Homo sapiens

Maybe the most concise definition of ROA appears in Jobling, Hurles & Tyler-Smith (2004):

> [...] the 'out of Africa' model proposes that the transition [from *erectus* to *sapiens*] took place recently (< 200 KYA) in Africa, and that these humans replaced the hominids already present on the other continents (Jobling, Hurles & Tyler-Smith, 2004:248),

but equivalent definitions can be found elsewhere, as, for example, Lewin (1998):

> The single, recent origin model, in which Africa serves as the source of modern humans, who then replaced established populations (Lewin, 1998:388, Figure caption).

In order to properly understand ROA, we need to discuss first its context, *both* historical (emergence, elaboration and trajectory towards (general) acceptance), and palaeoanthropological (the events happening before the putative threshold of modernity some 100-200kya, in Africa and elsewhere).

2.1.1. The historical development of ROA

Human evolution is, by its very nature, a historical endeavor and the theories it produces fundamentally involve time. Because time is conceived as linear, at least in the modern Western and westernized world (Eliade, 1981), it is extremely tempting for these theories to also become linear, describing a series of events strung on an imaginary axis. Even outside the popular press, there is a tendency to project "grand themes" onto such linear stories, usually, the myth of the hero (McBrearty & Brooks, 2000; Lewin, 1998:14-15), whereby the linearized story becomes teleological, a tale of purposeful change, from ape to man. Everything falls into place, like in any good novel (Leder *et al.*, 2002), when the hero (a pre-hominid ape) starts its long and tortuous journey towards full modernity, fighting evils (ice ages, droughts, predators, diseases, wars, cheaters), making friends (evolution of social life), discovering deeply hidden truths (a thrusting projectile can break bone, caring for one's slow-growing babies insures their survival), inventing untold devices and implements (bone spear tips, hand axes, spoken language) and slowly transforming into something better, something superior, the modern human as you and me, but also punishing those who failed

and stagnated somewhere in between (replacement of the locally evolved archaics). A tale where the hero is not a creature of flesh and blood, nor a mighty god, but a *concept*, a genetic lineage, an immaterial thing moving through time and inhabiting body after body, each time different, each time better.

Is this the way things *really* happened? Of course, the answer to such a question is '*we don't know!*'. But most probably not. Life as we know it, and the data gathered from a multitude of sources, strongly suggest that things are never so simple, that histories are rarely linear, and that myths almost always transmit what the hearer already knew (Griffiths & Kalish, 2005). However, modern scientific accounts of human evolution are *not* literary creations, by any account. But the change in language and techniques does not necessarily mean a change in the underlying ideational framework. The hero is still there, the myth still guides our search.

It seems very plausible that, by using general principles of folk-biology (Atran, 1998), people have tried to allocate themselves a place in the living world since the remotest prehistory. One of the earliest documented attempts at linking humans and other living primates were the anatomical parallels drawn between gladiators and Old World Monkeys by the Greek (living in Rome) physician Galenos of Pergamum[2]. Later, in his seventeenth century monumental work *Systema Naturae* (1735, updated throughout his entire life), Carl Linnaeus placed humans (*Homo diurnus* and later *Homo sapiens*) into the genus *Homo*, besides chimpanzees (*Homo nocturnus troglodytes*). Charles Darwin, in the *Descent of Man and Selection in Relation to Sex* (1871), argues that humans should not be allocated a genus of their own (Jobling, Hurles & Tyler-Smith, 2004:205).

In 1935, Louis Leakey published *The Stone Age Races of Kenya* (Leakey, 1935), where he, for the first time, articulates a primitive version of ROA, whereby modern humans emerged in Africa and later spread through the Old World, displacing the local archaic humans. Unfortunately, the archaeological basis for this claim, the robust modern fossils from Kanjera and the mandible from Kanam, Kenya, were subsequently shown to be much more recent (in the first case), probably of Holocene[3] age, and affected by pathology, in the second case (Trinkaus & Zilhão, 2003: 499). During the late seventies and early eighties, the idea of a recent African origin of moderns and their later dispersal was seriously considered in the

2 http://www.udayton.edu/~hume/Galen/galen.htm September 2006.

3 The *Holocene* represents the most recent geological epoch, starting only 10 kya and following the Pleistocene (Wilson *et al.*, 2000: 3-4)

light of the new archaeological evidence. This involved mainly the site of Border Cave (Lembombo Mountains, on the border between South Africa and Swaziland) and Europe, based on the proportions of the limb segments in fossil hominids (e.g., Beaumont, Villiers & Vogel (1978) for African finds and interpretations, Trinkaus (1981) for European Neanderthal versus modern human limb proportions, and Bräuer (1984) for morphological (cranium) approaches to modern human origins in Africa). The main themes of this period were that the earliest modern fossils (plus a few transitional forms) were found in Africa, and that Eurasia seemed to show an invasion of modern humans from some external location(s), most probably Africa.

Also, the eighties saw the appearance of a new and extremely important player: genetics. The first notable application of this exponentially-growing field to modern human origins, came to many as a surprise, as it used *living* people to make inferences about *extinct* hominids (Relethford, 2001). *Mitochondrial DNA and human evolution* by Rebecca Cann, Mark Stoneking and Allan Wilson, published in January 1987 in *Nature* (Cann, Stoneking & Wilson, 1987), analyzed mtDNA from 147 people originating from 5 geographic populations: 20 Africans (2 of Sub-Saharan origin, 18 Afro-Americans), 34 Asians (China, Vietnam, Laos, the Philippines, Indonesia and Tonga), 46 Caucasians (Europe, North Africa, and the Middle East), 21 aboriginal Australians and 26 aboriginal New Guineans. They inferred that Africa is the "likely source of the human mitochondrial gene pool" (Cann, Stoneking & Wilson, 1987:33) because "one of the two primary branches leads exclusively to African mtDNAs [...] while the second primary branch also leads to African mtDNAs" (Cann, Stoneking & Wilson, 1987:33). They observe that "each non-African population has multiple origins [and] each area was colonised repeatedly" (Cann, Stoneking & Wilson, 1987:33 and 31).

They also attempt to attach a timescale to this phylogenetic tree, by assuming a molecular clock, and arrive at an estimated age of the MRCA[4] of all living mtDNA lineages of 140-290ky and a migration out of Africa (dated by the MRCA of all mtDNA lineages not containing African branches) around 90-180kya. The non-existence of "extremely divergent types of mtDNA in present-day Asians, more divergent than any mtDNA found in Africa" (Cann, Stoneking & Wilson, 1987:35) lead them to "propose that *Homo erectus* in Asia was

4 Most Recent Common Ancestor.

replaced without much mixing with the invading *Homo sapiens* from Africa." (Cann, Stoneking & Wilson, 1987:35-36), a conclusion apparently supported by their reading of the archaeological record and nuclear DNA studies (Cann, Stoneking & Wilson, 1987:35). One bold interpretation made in the paper's abstract, that "All these mitochondrial DNAs stem from one woman who is postulated to have lived about 200,000 years ago, probably in Africa" (Cann, Stoneking & Wilson, 1987:31), will ignite the popular imagination, energetically supported by publications like the second part of Bryan Sykes *The Seven Daughters of Eve* (Sykes, 2004), continued by his private business, *Oxford Ancestors* (www.oxfordancestors.com), where, simply by donating your mtDNA, you will be told the name of the specific daughter of Eve being your "mitochondrial mother": Ursula, Xenia, Helena, Velda, Tara, Katrine, Jasmine or Ulrike (and, oh, it works only if you are European).

2.1.1.1. Mitochondrial DNA

Cann, Stoneking & Wilson (1987) represented a turning point in the history of ROA, as it was the first genetically-based study to explicitly support this model. Nevertheless, there were a number of critiques (see Relethford, 2001 and Jobling, Hurles & Tyler-Smith, 2004) summarized here, but, overall, later and better designed and controlled studies confirmed the basic finding, namely, that mtDNA has a fairly recent MRCA located in Africa.

The *mitochondria* are cell organelles essential for energy production (Seeley, Stephens & Tate, 2005:86). They are responsible for the oxidative metabolism and most ATP synthesis (Seeley, Stephens & Tate, 2005:86). There is a variable number of copies of mitochondria per cell (hundreds to thousands, Jobling, Hurles & Tyler-Smith, 2004:39), depending mainly on the cell's energetic requirements, and which can be increased by division of preexisting mitochondria (Seeley, Stephens & Tate, 2005:86). Evolutionarily, the mitochondrion originated as a free-living ancestral prokaryote[5], most closely related to modern-day α-proteobacteria[6] (Kutschera & Niklas, 2005:7). Somewhere around 2.2-1.5 bya, these free-

5 A *prokaryote* is a cell devoid of a nucleus, as opposed to *eukaryotes*. In the first case, the cell's genome is a closed double-stranded DNA loop, contained directly by the cytoplasm, while in the second case, the DNA is structured in a number of chromosomes, contained in the nucleus (Skelton, 1993:81).

6 More specifically, *Rickettsia*, an obligate intracellular parasite causing typhus (*R. prowazekii*, Lewin, 2004:78) as well as some other pathologies (Rocky Mountain spotted fever - *R. rickettsii*, Rickettsial pox - *R. akari*, etc.; see http://microbewiki.kenyon.edu/index.php/Rickettsia September 2006).

living ancestors were taken up by a host cell (it is still debated if this had a genomic organization already eukaryotic or still prokaryotic, Arachaebacteria-like) and through co-evolution (Skelton, 1993:52-55), they became endosymbionts. This inter-relationship developed so far that, currently, the symbiosis is obligate and most of the mitochondrial genome was transferred to the nucleus.

This *endosymbiotic theory* (Skelton, 1993:896; Maynard-Smith & Szathmáry, 1995:137; Jobling, Hurles & Tyler-Smith, 2004:41; Kutschera & Niklas, 2005:2; Lewin, 2004:78), has a very long history, beginning in the early twentieth century with a theoretical paper (Mereschkowsky, 1905) arguing for the "xenogenous origin of organelles" (Kutschera & Niklas, 2005:5). The idea was mostly forgotten, being considered either too speculative or utterly wrong (Kutschera & Niklas, 2005:6), until Lynn Margulis revived it in the late sixties (Margulis & Sagan, 1997). The theory gained momentum and, currently, it is generally accepted and used as a textbook example of biological evolution (Skelton, 1993:894). It has also been extended to other organelles [chloroplasts (Margulis & Sagan, 1997; Kutschera & Niklas, 2005; Skelton, 1993:899; Lewin, 2004:78-79), nucleus (Kutschera & Niklas, 2005:12-13; Maynard-Smith & Szathmáry, 1995:136; Skelton, 1993:899) and microtubules (Maynard-Smith & Szathmáry, 1995:142)].

The mitochondrion still contains its own genome, even after transferring its largest part to the separate cell's nuclear genome (transfer still continuing today[7]). These DNA sequences in the nuclear genome recognizably of mitochondrial origin are called *numts* (nuclear mtDNA insertions) and an early survey revealed that there are over 400 kb, almost 25 times as much DNA as that still contained in the mitochondrion itself (Jobling, Hurles & Tyler-Smith, 2004:41). Some of these *numts* are quite ancient, but there are some very recent, polymorphic even in human populations (Jobling, Hurles & Tyler-Smith, 2004:41). The mtDNA is organized as a circular double-strand molecule (Jobling, Hurles & Tyler-Smith, 2004:40), 16,596 bp long in humans (Lewin, 2004:78), and is extremely compact, from an informational point of view: "there are no introns[8], some genes overlap and almost every single base pair can be assigned to a gene" (Lewin, 2004:78).

---

7 At an estimated rate of between $2 \cdot 10^{-5}$ and less than 10^{-10} per generation (Lewin, 2004:79).

8 An *intron* or *intervening sequence* represents a "segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (*exons*) on either side of it" (Lewin, 2004:991).

The exception is represented by the *D-loop* (or the *control region*, Jobling, Hurles & Tyler-Smith, 2004:61), which is involved in the initiation of DNA replication (Lewin, 2004:78) and does not encode proteins or RNAs (Lewin, 2004:61). There are 13 protein-coding regions and all proteins are necessary for respiration (e.g, cytochrome *b* and a unit of the ATPase, Lewin, 2004:78). There are also 22 tRNA and 2 rRNA genes (Lewin, 2004:78), necessary as the mitochondria have an idiosyncratic genetic code, different from the nuclear one[9] (Jobling, Hurles & Tyler-Smith,2004:41). The mutation rate of the mtDNA is much higher than for nuclear DNA (one order of magnitude, Lewin, 2004:60) and is not uniform along the molecule: lower in the coding regions and much higher in the control region, which contains two *hypervariable segments* (or *hypervariable regions*), HVR I and HVR II, where the mutation rate is so high that mutations can often be observed across a small number of generations (Jobling, Hurles & Tyler-Smith,2004:61).

The mitochondria have one property which makes them extremely interesting for human genetic studies: they are (almost) exclusively transmitted through the female line (Jobling, Hurles & Tyler-Smith, 2004:40; Seeley, Stephens & Tate, 2005:87; Lewin, 2004:75): the oocyte contains ~100,000 mitochondria while the sperm has only ~50-75 (required for motility), and there seems to exist an evolved mechanism selectively eliminating the paternal mitochondria from the egg[10] (Hurst & Werren, 2001; Jobling, Hurles & Tyler-Smith, 2004:40; Lewin, 2004:76), encoded by nuclear genes (Schwartz & Vissing, 2002). There are cases of paternal inheritance of mitochondria in humans (Schwartz & Vissing, 2002; Kraytsberg *et al.*, 2004) and other animals (Rokas, Ladoukakis & Zouros, 2003; Tsaousis *et al.*, 2005), but its prevalence is still debated (Tsaousis *et al.* (2005) suggest high levels of recombination in animals). This could represent a potential problem for phylogenetic studies assuming clonal, maternal-only, transmission of mitochondria (Rokas, Ladoukakis & Zouros, 2003; Tsaousis *et al.*, 2005; Jobling, Hurles & Tyler-Smith, 2004:40,42-43), but the opinions range from negligible (Awadalla, 2004; Lewin, 2004:76), to important (Rokas, Ladoukakis & Zouros, 2003), to highly relevant (Kraytsberg *et al.*, 2004). Given that recombination between mitochondrial genomes (and *heteroplasmy*) is relatively frequent (Rokas, Ladoukakis & Zouros, 2003; Tsaousis *et al.*, 2005), but is different from paternal

9 The differences are in the allocation of one nuclear STOP codon to Tryptophane (UGA), two Arginine (AGA, AGG) codons to STOP and two Isoleucine (AUA, AUU) codons to Methionine (Jobling, Hurles & Tyler-Smith, 2004:41; Lewin, 2004:175 also provides a phylogeny of these changes).

10 Such a mechanism prevents selfish evolution of mitochondria, whereby mutants detrimental to the host but better able at invading the egg would be naturally selected (Hurst & Werren, 2001).

inheritance (Awadalla, 2004), which seems rare (Schwartz & Vissing, 2002; Awadalla, 2004), at least in humans, and that phylogenetic studies using mtDNA are intrinsically probabilistic, I do not think that paternal inheritance of mitochondria would radically transform the interpretation of human mtDNA studies, but further data will allow a proper evaluation of its impact. For the moment, thus, I agree with a negligible to moderate position (Awadalla, 2004; Lewin, 2004:76).

Annex 1 defines the concepts of the *Most Recent Common Ancestor* (MRCA) for non-recombining DNA lineages (focusing on mtDNA), *coalescence theory*, *gene genealogy* and *expected coalescence time*. When applied to Cann, Stoneking & Wilson's (1987) data, the coalescent theory gives a 95% CI of 152-473kya (with mean 290kya) (Templeton, 1993:58), highlighting the potential problems with the molecular clock.

2.1.1.2. The Molecular Clock

The main concept behind Cann, Stoneking & Wilson's (1987) claim concerning the age of mitochondrial Eve, is the *molecular clock*. The idea has a long history, starting in the sixties with E. Zuckerkandl and L. Pauling, who suggested (Zuckerkandl & Pauling, 1962, 1965) that the rate of change of aminoacids in proteins was approximately constant over evolutionary time and across taxa. Later, V. Sarich and A. Wilson (Sarich & Wilson, 1967) studied the immunological distances (Sarich & Wilson, 1967:1200) between humans, chimps (both species), gorilla, orang-utan, siamang, gibbon and six species of OWM (Old World Monkeys) (Sarich & Wilson, 1967:1201), and, by assuming a 30mya age-of-split between OWM and hominoids on palaeontological grounds plus a linear relationship between evolutionary time and immunological distance (Sarich & Wilson, 1967:1202), they derived a 5mya split between humans and the rest of the African apes (Sarich & Wilson, 1967:1202, and the phylogenetic tree in Figure 1, p. 1201).

In its modern form, the molecular clock hypothesis rests on Motoo Kimura's *neutral theory of molecular evolution* (Kimura, 1968, 1983; Jobling, Hurles & Tyler-Smith, 2004:147; Halliburton, 2004:370-376), which states that the majority of molecular variation in living organisms is selectively neutral and its frequency in the population is controlled by genetic drift alone (Halliburton, 2004:370; Jobling, Hurles & Tyler-Smith, 2004:147). Given that

neutral mutations are random events with low probability[11] (Jobling, Hurles & Tyler-Smith, 2004:147; Halliburton, 2004:183-217), one can assume that their rate, for any given DNA sequence, is more or less constant across evolutionary lineages (Jobling, Hurles & Tyler-Smith, 2004:147; Halliburton, 2004:378). To be applicable for dating, any molecular clock must be *calibrated*, not unlike many other clocks, using an external dating method (Jobling, Hurles & Tyler-Smith, 2004:148).

But even if the molecular clock hypothesis is widely used, it has been repeatedly criticized. *Lineage effects* are those cases where intracellular processes impact on the mutation rate in different evolutionary lineages (Jobling, Hurles & Tyler-Smith, 2004:148):

- *generation time effect* assumes that because most mutations are produced during DNA replication in the germ line, and different evolutionary lineages have roughly the same number of cell divisions per generation, the mutation rate depends on generation length (Jobling, Hurles & Tyler-Smith, 2004:148). This seems supported by data in mammals (Jobling, Hurles & Tyler-Smith, 2004:148-149) and *Drosophila* (Halliburton, 2004:382). As the germline cells in males undergo much more divisions than in females, they tend to accumulate more mutations as the male's age increases, conducing to the *male-driven evolution* hypothesis (Jobling, Hurles & Tyler-Smith, 2004:149);
- *metabolic rate hypothesis* assumes that most mutations are induced by endogenous mutagens (primarily free radicals; Jobling, Hurles & Tyler-Smith, 2004:149), implying that organisms with higher metabolic rates should also present more elevated mutation rates;
- *DNA repair*: the efficiency of DNA repair mechanisms and/or the capacity of the cellular processes to neutralize mutagens before they disrupt the DNA sequence could be different in different lineages (Jobling, Hurles & Tyler-Smith, 2004:149).

There are also *cross-lineages inconsistencies*, like:

- *rate difference between nonsynonymous and synonymous substitutions*: a synonymous mutation does not alter the aminoacid sequence of the resulting protein, while a nonsynonymous mutation does (Jobling, Hurles & Tyler-Smith, 2004:56). Two competing explanations have been proposed: that short pulses of selection at

11 Some estimates are in Halliburton (2004:194-195).

nonsynonymous sites distort the molecular clock (Gillespie, 1984:8011) and the other based on T. Ohta's *near neutral theory* (Halliburton, 2004: 388-394; Ohta, 1996), which considers not only neutral mutations, but also slightly deleterious ones (nonsynonymous substitutions): these behave as neutral in small populations and as deleterious (i.e., selected against) in large populations.

2.1.1.3. How neutral really is the human mtDNA?

Also important is the assumption of neutrality of the mtDNA control region. This assumption is essential for its phylogenetic usage (Jobling, Hurles & Tyler-Smith, 2004:154-195), as, for example, an undetected selective pressure can be taken as a sign of population expansion. During the last years, a series of studies appeared, questioning the neutrality of the mtDNA and presenting cases of selective pressure on this molecule. For example, Howell *et al.* (2004) show that the coding region of certain lineages, even if evolving non-neutrally, is a good approximation of a molecular clock, while the neutral control region behaves as a molecular clock only at the haplogroup level, implying that the two have evolved separately from one another.

On the other hand, Moilanen *et al.* (2003) found that not only is there selective pressure on mtDNA, but that this pressure differs between phylogenetic lineages and mtDNA regions. Mishmar *et al.* (2003) tried to identify the specific selective pressures and showed that, while the African mtDNA does not deviate from neutrality (Mishmar *et al.*, 2003:172), the European, Asian, Siberian and Native American lineages do (Mishmar *et al.*, 2003:172). The most divergent mtDNA gene is *ATP6*[12] (Mishmar *et al.*, 2003:174), and the authors' suggestion is that this is due primarily to climate and diet-related selection (Mishmar *et al.*, 2003:176), with a large impact on phylogenetic studies using mtDNA; a recent paper by Mau, Lee & Tzen (2005) does confirm the actual locus of selection as being the *ATP6* gene[13]. The existence of pathologies determined by mtDNA mutations is another hint that there might have been selective pressures shaping the distribution of mtDNA polymorphisms in human populations (Jobling, Hurles & Tyler-Smith, 2004:254).

12 One of the most conserved mtDNA genes in the animal kingdom.
13 Probably 25 human-specific aminoacid residues (Mau, Lee & Tzen, 2005:146).

If these claims about the non-neutrality of mitochondrial DNA turn out to be valid, then some inferences based on this molecule will have to be reevaluated, including the time to the MRCA (Halliburton 2004:462) and the interpretation of mtDNA branches as selective sweeps as opposed to demographic events.

2.1.1.4. Review of critiques of Cann, Stoneking & Wilson (1987)

The paper was criticized almost immediately, concerning the methods, assumptions and data (Relethford, 2001:78; Jobling, Hurles & Tyler-Smith, 2004:255; Templeton, 1993):

- *the sampling*: out of 20 Africans, only 2 were of Sub-Saharan origins, the others being Afro-Americans. This raises the legitimate problem of admixture in the Americas, mainly with European lineages, but, on historical grounds, this seems to be a rather minor possibility as it would involve sizable admixture between women of European descent and men of African ancestry. Nevertheless, this sampling procedure could bias the results, especially the estimated age to the MRCA (Relethford, 2001:78; Jobling, Hurles & Tyler-Smith, 2004:255);
- *the mutation rate*: this is a fundamental parameter, as a smaller value (faster mutation) would produce a younger MRCA (Relethford, 2001:78). The authors used an estimated mtDNA mutation rate of 2-4% per million years (Cann, Stoneking & Wilson, 1987:33), derived from comparing the degree of differentiation between humans in Australia, New Guinea and the Americas with their colonization dates (Cann, Stoneking & Wilson, 1987:33), as well as from studies of other species (Cann, Stoneking & Wilson, 1987:34). But this assumption of constancy can be attacked (Section 2.1.1.2);
- *the tree inference*: the method used is thoroughly analyzed by Templeton (1993:51-54), and the main problem is that the tree reported as the most parsimonious by the authors proves not to be so (Templeton, 1993:52). The alternatives found by Maddison (1991:358) numbered no less than 10,000 trees of length 307, 5 steps shorter than the supposed most parsimonious one (Cann, Stoneking & Wilson, 1987:34) and, more than that, these alternatives featured more than 50% "mixed-sister trees" (Maddison, 1991:358), having a mixed Asian-African clade. This resulted, ultimately, from a non-optimal usage of the tree inference package PAUP[14],

14 http://paup.csit.fsu.edu/

which was sensitive to the input order of the taxa when being attracted into a local optimum, thus, requiring several runs with randomized input order (Templeton, 1993:52-53);

- *the root location*: the algorithm used to produce a rooted tree out of the unrooted one generated by PAUP was "by placing the root [...] at the midpoint of the longest path connecting the two lineages" (Cann, Stoneking & Wilson, 1987:34, Figure's 3 caption). This could be biased towards an African origin if, for example, the rate of mutation accumulation was higher in Africa than elsewhere (Relethford, 2001:78; Jobling, Hurles & Tyler-Smith, 2004:255) and the rooting of the tree using an *outgroup* is to be preferred (Jobling, Hurles & Tyler-Smith, 2004:255).

2.1.1.5. Reaching the current view on modern human mtDNA

Answering the critiques summarized in Section 2.1.1.4 above, a paper appeared in 1991 in *Science* (Vigilant *et al.*, 1991), where a sample of 189 individuals was studied[15], including 121 native Sub-Saharan Africans (Vigilant *et al.*, 1991:1503), and the resulting tree was rooted using as outgroup a chimpanzee sequence (Vigilant *et al.*, 1991:1504). The obtained tree (Vigilant *et al.*, 1991:1505, Figure 3) was similar to the original Cann, Stoneking & Wilson (1987:34, Figure 3), in featuring a primary division between an African-only and a mixed African-non-African group, and an estimated age to the MRCA of human mtDNA of 166-249ky (mean 208kya) (Vigilant *et al.*, 1991:1506), in very good agreement to the earlier Cann, Stoneking & Wilson's (1987:34) 140-290kya.

Most of the later studies using mtDNA have confirmed these basic findings and the current consensus today is that human mtDNA features (Jobling, Hurles & Tyler-Smith, 2004:252; Relethford, 2001:91):

- a TMRCA of 172±50kya and a MRCA of the mixed African-non-African branch of 52±28kya;
- deep branches within Africa and star-like branches outside Africa;
- complete separation of African and non-African lineages.

All these, taken together, could be taken to suggest that mtDNA supports a ROA model, with an estimated origin of all living human mtDNA lineages in Africa ~170kya, followed by

15 By sequencing 1122 bp of the control region (Vigilant *et al.*, 1991:1504) as opposed to the earlier Cann, Stoneking & Wilson (1987:32) restriction mapping.

migration out of Africa and expansion of a subset of the African lineages, but, we must bear in mind all the possible caveats discussed in the previous sections. As always in science, an apparent consensus does not necessarily imply truthfulness, and today's consensus will possibly be tomorrow's folly.

2.1.1.6. ROA in its current form

In March 1988, long before the problems with Cann, Stoneking & Wilson (1987) were clarified by Vigilant *et al.* (1991) and subsequent work, and while the popular press was inflamed by the modern story of the African Garden of Eden (e.g., the January 11, 1988 issue of Newsweek[16]), Chris Stringer and Peter Andrews published a paper, "Genetic and Fossil Evidence for the Origin of Modern Humans" in *Science* (Stringer & Andrews, 1988). They compared two competing models, acknowledged by the authors to be "extreme" (Stringer & Andrews, 1988:1263), but whose comparison "should allow the clearer tests for the models from existing data, tests which are not feasible for several other proposed models." (Stringer & Andrews, 1988:1263).

These models are the "regional continuity (multiregional origins)" and "Noah's Ark (single origin)" (Stringer & Andrews, 1988:1263). Briefly, the multiregional model considered by Stringer & Andrews (1988) emphasizes the regional continuity and evolution *in situ* of modern humans, mediated by global gene flow, while their "Noah's Ark" is described as:

> the single origins model assumes that there was a relatively recent common ancestral population for *Homo sapiens* which already displayed most of the anatomical characters shared by living people [...] proposed Africa as the probable continent of origin of *Homo sapiens*, with an origin of the species during the late Pleistocene[17], followed by an initiation of African regional differentiation, subsequent radiation from Africa, and final establishment of modern regional characteristics outside Africa (Stringer & Andrews, 1988:1263).

This represents the first explicit definition of the modern ROA, as it is currently understood in the literature (Trinkaus & Zilhão, 2003:499, but see Wolpoff & Caspari, 1997 for the long history of this idea):

- the ancestral population for modern humans is *recent* and understood to be *unique*,

16 The cover story's "The Search for Adam and Eve" author, Karen Springen, was awarded the 1988 AAAS Westinghouse Science Writing Award.

17 The Pleistocene (1,750-10kya) immediately precedes the Holocene and follows the Pliocene (Wilson *et al.*, 2000:4).

opening thus the door for discussions of the speciation of *Homo sapiens* in an isolated population;

- this population already had the (almost complete) constellation of anatomical modern features;
- it was located in Africa;
- initial African differentiation and subsequent dispersal with replacement.

The authors support their views with palaeoanthropological data, highlighting that convincing transitional fossil forms were abundant in Africa and absent elsewhere and that, outside Africa, the transition from archaic *Homo* to its modern forms was abrupt, and also with genetic data, an important place being accorded to Cann, Stoneking & Wilson (1987).

During the following years, most findings, archaeological, palaeoanthropological and genetic, have supported this view, so that it became established as *the default* theory of modern human origins, largely regarded as *true* and not as a hypothesis anymore. So that, for example, Stephen Oppenheimer claims that: "[...] the original out-of-Africa picture suggested by the mitochondrial markers has emerged triumphant, and the multiregionalists have become an isolated, albeit vociferous minority." (Oppenheimer, 2004:50).

But who were these *multiregionalists*, so much ridiculed by Oppenheimer and many others, apparently defeated in their views but still reluctant to let go? And what *regional continuity* are they talking about? Continuity of *what*?

2.1.2. Before the moderns: the palaeoanthropological context

It is universally accepted that modern humans are just another species of *Mammals*[18]. More exactly, we belong into the order *Primates*, which we share with some other 365 living species (Groves, 2001). This order is rather vaguely defined morphologically (Jobling, Hurles & Tyler-Smith, 2004:204) but there is a series of synapomorphies[19], like binocular

18 Mammals are defined as endothermic vertebrates with mammary glands, producing milk in the females, with hair or fur and a four-chambered hart; their classification is still evolving (Springer *et al.*, 2004; McKenna & Bell, 1997).

19 A *synapomorhpy* is a derived character shared by the members of a clade to the exclusion of the other forms from which it diverged (Skelton, 1993:528-529). A *clade* is defined as all the descendants (and only them) of an ancestral form, plus this ancestor (Skelton, 1993:518).

vision and shortened muzzle or snout (Jobling, Hurles & Tyler-Smith, 2004:204). A phylogenetic tree of *Primates* is reproduced Figure 1 (modified from Jobling, Hurles & Tyler-Smith, 2004:204[20] and Groves, 2001). The phylogeny itself is not much debated, but the dates of the splits are (Jobling, Hurles & Tyler-Smith, 2004:217, 218-219: Box 7.7); I will use Raaum *et al.* (2005), which analyzed the entire mtDNA genomes of selected *Catarrhines* , with strict criteria defined for the calibration points from the fossil record and a penalized likelihood method[21] (Sanderson, 2002). The age of the entire *Primates* order was estimated to be ~80-90mya, and the age of the most recent common ancestor of chimpanzees and humans was estimated to have lived ~6mya (but see Section 2.2.4).

The ancestor of the modern anthropoids (*Simiiformes*) is estimated to have lived ~30-40mya, and the standard theory assumes an African origin, but some recent finds in Bugti Hills, Pakistan (Marivaux *et al.*, 2005; Jaeger *et al.*, 1998), seem to challenge this view (Jaeger & Marivaux, 2005). They consider, instead, an Asian origin and, possibly, early evolution, of anthropoids with a later change of focus on Africa, suggesting a much more complex evolutionary history, and, probably, Asian-African exchanges (Jaeger & Marivaux, 2005; Dawkins, 2004:117-123).

---

20 Their phylogenetic tree, page 204, contains a slight error, as the OWM plus *Hominoidea* form the *Catarrhini* and the NWM the *Platyrrhini*, and not the reverse.

21 This method can accommodate the differences in evolutionary rates between lineages, their phylogeny and multiple calibration points, and is implemented by the program `r8s`, available at http://ginger.ucdavis.edu/r8s/ September 2006.

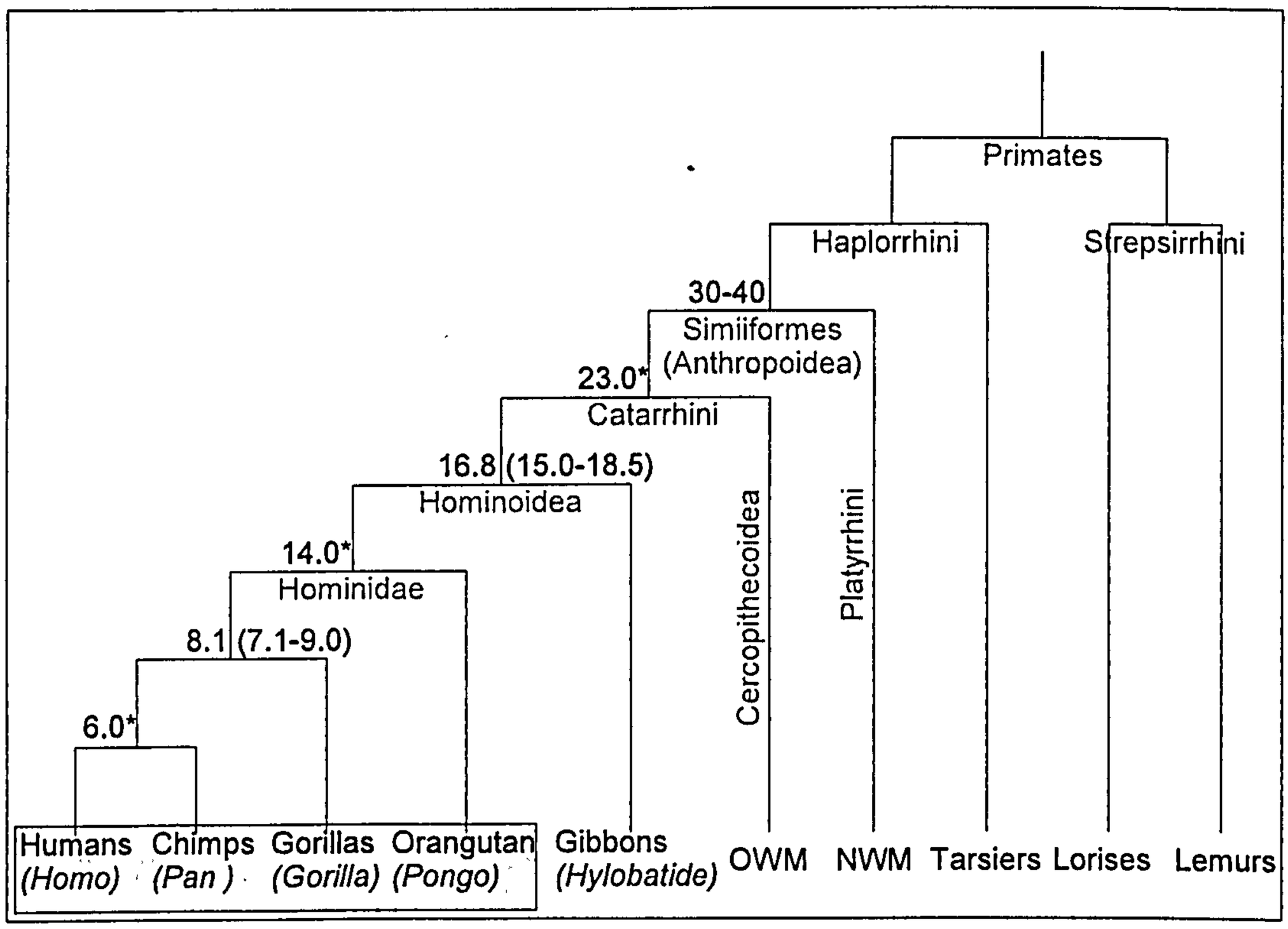

Figure 1: The phylogeny of living primates.

Terminal leaves: species or entire clades; OWM = Old World Monkeys, NWM = New World Monkeys; Boxed area (humans, chimps, gorillas and orangutans): the Hominidaes. *The split dates marked with (*) are derived from the fossil record, the others are based on the entire mtDNA genome, with their 95% confidence intervals (Raaum et al., 2005:252), measured in million years before present (mya).*

A note on terminology is necessary, as at least three different concepts appear in the literature: *hominoid*, *hominid* and *hominin*: their meaning depends mostly on the implicit phylogenetic assumptions of the author. In accordance with what seems to be the most widespread modern usage[22] (Antón, 2003; Raaum *et al.*, 2005; Leigh, 2006; Stefansson *et al.*, 2005; Strait & Grine, 2004; Goren-Inbar *et al.*, 2004; Trauth *et al.*, 2005; Falk *et al.*, 2005; Bobe & Behrensmeyer, 2004; Garrigan *et al.*, 2005a, b; Cameron, 2003; Holliday, 2003; Weber, Czarnetzki & Pusch, 2005; Villmoare, 2005), and agreeing partially with Begun (2004:1480, note 2) and implicitly with Jobling, Hurles & Tyler-Smith (2004:204), I will mean by *hominoids* all extant or extinct primates related to humans and great apes (the clade including humans, chimps, gorillas, orangutans and gibbons), by *hominids* its sub-clade including only humans, chimps, gorillas and orangutans, while *hominins* will refer to

22 In the older literature (but not only), *hominids* are understood to refer either to modern *hominids* or *hominins*, and usually the context is sufficient to disambiguate between them.

humans and the fossil hominids most closely related to them (see also Relethford, 2003:38-41).

The MRCA of modern humans and chimpanzees was most probably living in Africa ~6mya (however, see Section 2.2.4), but its exact identification with any of the known fossils is still debated. *Sahelanthropus tchadensis*[23] (Brunet *et al.*, 2002), a fossil hominid ~7my old, discovered in Chad, was judged to be a common ancestor[24] of humans and chimpanzees based on its cranium, face (including browridges) and teeth, but Brigitte Senut (2002, cited in Jobling, Hurles & Tyler-Smith, 2004:237) contends that it could represent merely a female (proto-)gorilla. This claim must be taken with a grain of salt (or two) as she is one of the discoverers of *Orrorin tugenensis*[25], and a supporter, with Yves Coppens, of the *East Side Story* (Coppens, 1991), proposing that the opening of the Rift Valley and the ensuing climatic changes in East Africa have prompted different evolutionary trajectories on its both sides, with only East African hominins evolving bipedality, bigger brains, etc[26]. *Orrorin tugenensis* (Senut *et al.*, 2001) dates to ~5.8-6.1mya and his probable upright walking (but possibly still a good climber) and small molars with thick enamel link it with some probability to the lineage leading to modern humans (Jobling, Hurles & Tyler-Smith, 2004:237; Senut *et al.*, 2001:142). A third contender to this special place is *Ardipithecus ramidus kadabba*[27], discovered at Middle Awash, Ethiopia (Haile-Selassie, 2001), dated to 5.2-5.8mya, and considered to belong to the evolutionary line leading to humans by Haile-Selassie (2001:180). There is no consensus reached so far and, it seems probable that until more complete skeletal remains will be uncovered, the relative place of these three hominids will remain debated. But all these early hominids seem to have shared the same size (chimp-like), upright walking (thus, a very early character of the lineage leading to us) and forested habitats (Jobling, Hurles & Tyler-Smith, 2004:237), picturing a coherent image of our earliest hominin ancestors.

---

23 Nicknamed *Toumai*, "hope of life" in the Dazaga language of Chad [dzg].

24 But probably not the last common ancestor, given the divergence between its age and the inferred age of split on genetic grounds.

25 Also known as "The Millennium Man", named after the Tugen Hills in Kenya (Senut *et al.*, 2001:138), the place of its discovery. "Orrorin" means "the original man" in the Tugen language [tuy] (Senut *et al.*, 2001:138).

26 Unfortunately for Senut's remark, Yves Coppens publicly accepted that his East Side Story theory can no longer be supported, given the discoveries of Toumai (*Sahelanthropus tchadensis*) and Abel (*Australopithecus bahrelghazali*) on the *west* side of the Rift Valley, in Chad, in *La Recherche* No. 361/2003.

27 "Kadabba" means "basal family ancestor" in the Afar language [aar] (Haile-Selassie, 2001:180).

Until the appearance of the genus *Homo*, some 2mya, Africa saw a radiation of different species of hominins (Cameron, 2003; Jobling, Hurles & Tyler-Smith, 2004:238-240; Strait & Grine, 2004), the most important being the *australopithecines*, which appeared ~4mya (Jobling, Hurles & Tyler-Smith, 2004:238), small-bodied, upright walking, highly sexually dimorphic primates with a small brain[28]. They suggest that bipedalism preceded any significant increase in brain size by some 2 million years, proving that something else than the freed hands, enlarged visual field or two-legged walking prompted this increase. They probably used and manufactured tools (Lewin, 1998:281), as suggested by the tool-use of modern chimps[29], implying a similar capacity in our last common ancestor, which predated the australopithecines (Jobling, Hurles & Tyler-Smith, 2004:246), but the detection of such archaeological assemblages is difficult given that they do not differ much from natural objects (Jobling, Hurles & Tyler-Smith, 2004:246; Panger *et al.*, 2002:239). There are many forms of *Australopithecus* (Jobling, Hurles & Tyler-Smith, 2004:238-240; Lewin, 1998:263-281; Cameron, 2003), some almost unanimously recognized as different species, some still debated (Cameron, 2003; Jobling, Hurles & Tyler-Smith, 2004:238-240; Strait & Grine, 2004), but it seems probable that *A. anamensis* or *A. afarensis* can be considered as our ancestors, while the others represent forms which left no modern descendants. In this same category seem to also fall the *Paranthropus*[30] and *Kenyanthropus*[31] genera (Jobling, Hurles & Tyler-Smith, 2004:239-240; Strait & Grine, 2004:438; Cameron, 2003).

The genus *Homo*, to which we also belong, seems to have appeared ~2mya in Africa and the first secure such fossils belong to *Homo erectus*[32]*/ergaster*[33] (Jobling, Hurles & Tyler-Smith, 2004:240; Lewin, 1998)[34]. It was characterized mainly by an enlarged brain (750-1225 cm^3, but possibly as low as 650 cm^3 for the Dmanisi D2282 specimen (Gabunia *et al.*, 2000:1022) or 600 cm^3 (Vekua *et al.*, 2002:88), overall anatomical similarity to modern humans, large body size, and indisputable tool use (Jobling, Hurles & Tyler-Smith, 2004:240-241, 247-

28 400-500 cm^3, proportionally the same size as for modern chimpanzees (Jobling, Hurles & Tyler-Smith, 2004:239; Lewin, 1998:275).

29 And other non-human primates (van Schaik, Deaner & Merrill, 1999; Panger *et al.*, 2002; Moura & Lee, 2004).

30 Heavily build hominids with large jaws and chewing teeth, small brains, probably adapted to low-calories diet (foliage, roots) (Jobling, Hurles & Tyler-Smith, 2004:239-240).

31 Described in Brunet *et al.* 2002:145-146.

32 "Upright man" in Latin.

33 "Working man" in Greek.

34 The older *Homo habilis* ("Skillful man" in Latin) seem to have been relegated to the status of *Australopithecus habilis* (Jobling, Hurles & Tyler-Smith, 2004:240).

248). It is the first hominin to spread out of **Africa** [its earliest remains, both skeletal and associated tool assemblages, in Africa are: **Koobi Fora**, 1.88-1.9mya (Antón, 2003:128); **Turkana Basin**, 1.8mya (Walker, 2002:39)], shortly after its appearance (Jobling, Hurles & Tyler-Smith, 2004:240-241; Dennell:2003:421-422). Its remains have been discovered in **Georgia** [**Dmanisi**, around 1.7-1.8mya (Gabounia *et al.*, 2002; de Lumley *et al.*, 2002; Gabunia *et al.*, 2000; Vekua *et al.*, 2002; Jobling, Hurles & Tyler-Smith, 2004:241)], **Indonesia** [**Trinil**, 1.0-0.7mya (Schwartz, 2004:53); **Sangiran**, 1.66±0.04mya (Dennell, 2003:430); **Mojokerto**, 1.8mya (Coqueugniot *et al.*, 2004:299)], **China** [**Xiaochangliang**, 1.36mya (Zhu *et al.*, 2001); **Gongwangling**, 1.15mya (Wang *et al.*, 1997:228), **Yunxian**, 0.8mya (Wu, 2003:132)] and **Israel** [**Erk-el-Ahmar**, 2.0-1.7mya (Ron & Levi, 2001); **'Ubeidiya**, 1.4mya (Antón, 2003:130)]. This geographic expansion seems to have followed the climatic events of the Pleistocene and is correlated with vegetal and faunal range dynamics (Dennell, 2003; Storm, 2001; Finlayson, 2005; Bergh, de Vos & Sondaar, 2001).

A very interesting twist to the story is given by R. Dennell and W. Roebroeks (2005), where the arguments for an African origin of *Homo* are analyzed and an alternative Asian origin is proposed. It is suggested that after the first exclusively African stages, (an) unspecified hominin(s) migrated into Asia (Dennell & Roebroeks, 2005:1100), where it developed into *Homo* (Dennell & Roebroeks, 2005:1101), migrating back to Africa. They also argue that "[i]t is not the continent that matters in studying human origins so much as the type(s) of environment with which early hominins were associated" (Dennell & Roebroeks, 2005:1102), namely, the *Savannahstan* - "the Pliocene grasslands extending from west Africa to north China" (Dennell & Roebroeks, 2005:1102). This theory seems very plausible and reminiscent of the models discussed for later stages of *Homo*, but, in the following, I will assume the (still) standard model.

This immense area where such early hominin presence is found raises two inter-related questions, one obvious and the other undeservingly neglected: how many species of early *Homo* were there and how continuous was their colonization? To the first question, the answers are:[35] one, two or many, while for the second, the answer depends on the specific region and time.

35 *Splitters* versus *lumpers* refer to an old debate in biology concerning the approach to classification: while splitters tend to see as many taxa as possible, lumpers prefer to group them together as much as possible (Holliday, 2003; Wolpoff & Caspari, 1997:67-68).

2.1.2.1. What are *species?*

Unfortunately, there are many problems with the immensely important concept of *species* throughout the biological sciences (West-Eberhard, 2003:526-563; Skelton, 1993:372-380; Howard & Berlocher, 1998:19-78; Hey, 2001; Tattersall & Mowbray, 2005; Holliday, 2003). It is generally agreed that "the biotic world is self-evidently 'packaged' into units" (Tattersall & Mowbray, 2005:371) and it is almost intuitively clear what a species should be. And yet, as Jody Hey (2001a, b) argues, species counts are generally meaningless because our cognitive biases force us to impose clear boundaries on an intrinsically messy world (Hey, 2001b:151ff). There is a plethora of proposed definitions for species[36] (Skelton, 1993:372-380; Howard & Berlocher, 1998:19-78; Hey, 2001; Tattersall & Mowbray, 2005; Holliday, 2003), each emphasizing a certain aspect of the process or its end-product, each having its own deficiencies (Skelton, 1993:372-380; Howard & Berlocher, 1998:19-78; Hey, 2001; Tattersall & Mowbray, 2005).

Biological species concept (BSC) or **Isolation Species Concept** (ISC): introduced and popularized by Ernst Mayr, highlights the reproductive coherence and distinctiveness of species. They represent "groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups" (Mayr, 1942:120; 1963:19). From a mathematical point of view, this is reminiscent of an *equivalence class* (Halmos, 2001), where the sexual reproduction is the equivalence relation dividing the living world into species. Unfortunately, despite its elegance and apparent intuitive appeal, it has many shortcomings, even for living organisms (Tattersall & Mowbray, 2005:373-374; Howard & Berlocher, 1998:22-23; Skelton, 1993:374-375):

- what does *potentially* mean? Are we supposed to experiment with allopatric populations?
- what is *reproductive isolation*? This implies a set of *isolating mechanisms* (Howard & Berlocher, 1998:22; Skelton, 1993:373), both *pre-* and *post-zygotic* (Skelton, 1993:373), such as *mechanical* (mechanically impossible mating), *sexual* (or ethological, involving decreased or absent mutual sexual attraction, due to different sexual display strategies), *ecological* (different ecological niches which do not intersect), *temporal* (different mating seasons or maturation schedules) and *gametic*

36 Jody Hey (2001b:327) lists 24.

(mating takes place, but the zygote fails to form), for the first type, and *hybrid inviability* (the embryo fails to develop or the individual dies before reaching maturity), *sterility* (the hybrid survives to maturity but fails to produce viable gametes) and *breakdown* (second-generation hybrids suffer reduced fitness) for the second, but there are frequent cases of imperfect application (Skelton, 1993:375; Howard & Berlocher, 1998:22-23; Section 2.2.4).

This concept is obviously inapplicable to asexual species (Skelton, 1993:375), and to extinct lineages (Skelton, 1993:374), or at least, not in a direct way.

Phylogenetic Species Concept (PSC): designed with morphology in mind by Joel Cracraft, defines a species as the "irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent" (Cracraft 1989, cited in Harrison, 1998:21). This highlights the well-known fact that "[...] speciation and morphological differentiation are the result of different, if potentially overlapping, sets of genetic processes" (Tattersall & Mowbray, 2005:375), allowing, thus, the recognition of morphological species across a potentially single biological species (Jolly, 2001:177).

Evolutionary Species Concept (ESC): due to George Gaylord Simpson, is defined as "a lineage (ancestor-descendant sequence of populations) evolving separately from others and with its own unitary evolutionary role and tendencies" (Simpson, 1961:153). It was designed with palaeontology in mind, but, unfortunately, it is too abstract (Tattersall & Mowbray, 2005:376) and neglects to specify the mechanisms by which species maintain their cohesion through time and how to actually identify how common evolutionary histories are to be assessed (Skelton, 1993:372).

Recognition Species Concept (RSC): emphasizes the *specific-mate recognition system*, which insures the reproductive cohesion of the species, defined thus as the "most inclusive population of individual biparental organisms which share a common fertilization system" (Paterson, 1985:15). Being a derivative of (and an attempt at correcting) the BSC (Skelton, 1993:375), it inherits most of its problems (Skelton, 1993:375), including the relative inapplicability to the fossil record (Tattersall &Mowbray 2005:375).

When dealing with fossil hominins, the task of identifying the species they belonged to is daunting (Skelton, 1993:461-486). This is because, on top of the usual complexities of species identification in living organisms, the investigator has to deal with a sparse and incomplete fossil record, where individuals are rarely found complete. The only reliable clues are frozen in the specimen's morphology[37] and the recognition of species based on morphology is extremely contentious (Tattersall & Mowbray, 2005), given that there is no direct, simple relationship between morphological differences and species status: there are different species with almost identical hard morphology and conspecific populations with marked morphological differences (Tattersall & Mowbray, 2005:374; Skelton, 1993:377-379). Moreover, there is also the problem of *chronospecies*, which represent two temporal stages in the history of the same lineage, which, "if [...] contemporaneous populations, we would have felt bound to recognize them as distinct species" (Skelton, 1993:464). When applied to the hominin fossil record, this raises the problem of the intrinsically subjective border between two consecutive chronospecies.

Addressing these issues, a number of heuristics have been proposed in the literature for delineating the hominin extinct species. One is to compare the morphological diversity (using either metrical or discrete characters) of a set of fossil individuals with that of extant model primates (usually the modern humans, the chimpanzees, or other great apes) (Tattersall & Mowbray, 2005:376-377; Cameron, 2003:3; Villmoare, 2005:684; Jolly, 2002), while another is to do a set-internal comparison and clustering[38] (the vast majority of palaeoanthropological studies). Given all these issues and the degree of subjectivity involved in the usage of heuristic methods, it is not surprising that there is a huge degree of controversy surrounding extinct hominin species (e.g., Wolpoff & Caspari, 1997:250-256).

I will follow Clifford Jolly's (2001) suggestion that the fossil hominins should be treated as a set of *allotaxa*, on the model of extant papionins (Section 2.2.4). Allotaxa are defined as "phylogenetically close, but well-differentiated and diagnosable, geographically replacing forms whose ranges do not overlap, but are either disjunct, adjoining or separated by comparatively narrow zones in which characters are clinally distributed" (Jolly, 2001:193-

37 Sometimes cultural markers are also used, as when different tool assemblages are used to infer the probable species of their maker.

38 Some authors apply cladistic principles, but this has been generally criticized as it implicitly assumes species-grade distinctions (e.g. Asfaw *et al.*, 2002:318).

194), implying that they belong to the same biological species (BSC) but possibly forming multiple phylogenetic species (PSC) (Holliday, 2003:657). A related concept is represented by the *syngameon*, originally defined in relation to a set of closely related species of plants which commonly hybridize (Lotsy, 1925, cited in Holliday, 2003:656): "plant taxonomists frequently group species in larger units called *syngameons*, within which natural hybridization may take place[, y]et the species within a syngameon remain separate species" (Skelton, 1993:375). Holliday (2003), building on Jolly (2001), argues that syngameons are much more common in the animal world than usually assumed.

Papio[39] and *Theropithecus*[40] are morphologically distinct (Jolly, 2001; Holliday, 2003:656), they have diverged ~5mya (Jolly, 2001:189), are usually classified as different genera (Holliday, 2003:656), and yet, they hybridize frequently under artificial settings (Jolly, 2001:189) and in nature (Holliday, 2003:657; Jolly, 2001:189), the resulting hybrids being viable and fertile (Holliday, 2003:657; Jolly, 2001:189-190, 197). Based on this and other such primate allotaxa, Jolly (2001), and especially Holliday (2003), conclude that the *hominins* were fully interfertile allotaxa during their entire existence through space and time: "[...] suggests that all human lineages stemming from the *H. ergaster*[41] stock were probably as fully interfertile as are extant *Papio* populations. On these grounds, they could be regarded as members of a single, polytypic (BSC) species" (Jolly, 2001:196) and that "a strict papionin analogy would therefore argue that all *Homo* (*sensu stricto*) were interfertile" (Holliday, 2003:659).

This does not assume, of course, a panmictic, homogeneous population of *Homo* throughout the Plio-Pleistocene Old World and does not imply the non-existence of regional characteristics and continuity. "[...T]he assumption of universal interfertility within the genus *Homo* (strictu sensu) [does not] conflict with evidence pointing to long-term, consistently diagnosable human lineages [...]" (Jolly, 2001:196).

39 Baboons, broadly distributed in Africa, composed of a debated number of species (Holliday, 2003:656-657).
40 Composed of a single extant species *(Theropithecus gelada*) circumscribed to the highlands of Ethiopia (Holliday, 2003:656).
41 Jolly uses *Homo ergaster* as the stem of all *hominins*, but for him this is just an allotaxa, not a biological species (Section 2.1.2.2).

2.1.2.2. Homo erectus and their feats

Homo erectus was born in the late '40s, when Ernst Mayr subsumed the previous taxa (*Pithecanthropus*, *Sinanthropus*, *Meganthropus* and *Telanthropus*) under a single name (Antón, 2003:126), but in 1983[42], during the Senckenberg conference, Chris Stringer, Peter Andrews and Bernard Wood proposed to split it into African and Asian species. This was backed by a proposed series of Asian autapomorphies and, because the Asian species also included the type specimen[43], it retained the *Homo erectus* name, while the African branch was named *Homo ergaster* (Kidder & Durband, 2004). The controversy started then continues today, but it seems that the balance is leaning towards a single encompassing, regionally variable, *Homo erectus* species (Jobling, Hurles & Tyler-Smith, 2004:240; Kidder & Durband, 2005:313; Gilbert, White & Asfaw, 2003:255; Asfaw *et al.*, 2002:319; Antón, 2003), even if this conclusion is not supported by all recent studies (Cameron, 2003; Villmoare, 2005; Schwartz, 2004). For example, one of the most important arguments in favor of a single widespread species is provided by a fossil discovered in the Dakanihylo member of the Bouri Formation, Middle Awash in Ethiopia and reported by Berhane Asfaw and colleagues (Asfaw *et al.*, 2002). This ~1my old specimen (brain capacity of about 995 cm^3, usually referred to as the "Daka cranium") clearly clusters with Asian *Homo erectus*, thus proving that "the early African and Eurasian fossil hominins represent demes of a widespread palaeospecies" (Asfaw *et al.*, 2002:317). Moreover, it represents an intermediate stage between earlier and later African specimens (Asfaw *et al.*, 2002:319), suggesting, overall, that "by 1Myr the taxon had colonized much of the Old World without speciating – a finding of considerable biogeographic and behavioural significance" (Asfaw *et al.*, 2002:319)[44].

Of course, as usually happens when subjectivity is coupled with fashion and funding policy[45], there is a plethora of proposed new "species" of fossil hominins. For example,

42 This account is mainly based on Kidder & Durband (2004:299-300).

43 For *Homo erectus*, the type specimen is considered Trinil 2, discovered by Eugene Dubois in 1891 in Java. See http://www.talkorigins.org/faqs/homs/typespec.html September 2006.

44 The methodology of the paper is later defended against various criticisms in Gilbert, White & Asfaw (2003) and the basic finding of a single *Homo erectus* species reiterated.

45 A find has to be "the find" in order to insure continuing funding of the project. See also the analysis of the media involvement in the *Homo floresiensis* case (Powledge, 2005). As Maciej Henneberg puts it: "[t]his abuse is especially tempting where individual researchers may gain professional standing through the creation of new categories. Discovering yet another fossil individual belonging to the human lineage is a great achievement, but an even greater one is discovering a whole new kind, a new ideal entity" (Henneberg, 2003:662).

Mallegni *et al.* (2003) rush to qualify the Ceprano finds as a new species "*Homo cepranensis*" on shaking grounds, to put it mildly (see the critique in Gilbert, White & Asfaw, 2003), while a recent Dmanisi hominin was attributed to the newly created "*Homo georgicus*" (Gabounia *et al.*, 2002) and assumed to represent an earlier stage than *Homo erectus* (but see Dennell & Roebroeks, 2005). Thus, I will assume, in the following, a single, polytypic, geographically widespread species named *Homo erectus*[46].

How good were the colonizing abilities of *Homo erectus*? Its earliest remains have been discovered around the Old World, but what does that mean? As discussed by Robin Dennell (Dennell, 2003) and hinted by others (e.g., Antón, 2003; Finlayson, 2005; Storm, 2001), there are three main questions concerning early hominin dispersals: *when* did it happen, *how often* and *how successful* were they? (Dennell, 2003:421). The third question, he argues, was usually neglected in the literature (Dennell, 2003:421), but is extremely relevant. It was usually implicitly assumed that the earliest date represents the beginning of a continuous colonization, an assumption hidden in depictions of phylogenies and maps (Dennell, 2003:422), but, in fact, it seems more plausible that these earliest events represent temporary occupations, heavily dependent on the climatic dynamics: "it may be more realistic to assume that they indicate a palimpsest of intermittent dispersal events, only some of which resulted in long-term colonization" (Dennell, 2003:422).

This view is supported by the sparsity of the fossil record, the ecological preferences of early *Homo erectus*, the competition with the other resident carnivores[47] and the increasing home-ranges towards northern latitudes (Dennell, 2003:422-424). This early population dynamics suggests that there were *core* areas of hominin occupation (the Rift Valley, the Levant), from which expansions, triggered by climatic events, ensued towards *peripheral* areas, intermittently occupied (Dennell, 2003:424). But as time passed by, the colonizing capabilities of *Homo* increased, due to biological and cultural changes, so that, "the Pleistocene record for hominids in the Old World can [...] be seen as a game between one side of increasingly proficient hominids, and another side of an increasingly disruptive climate[48]" (Dennell, 2003:424). This slowly changed, and only ~1mya the non-African

46 Compatible with what is sometimes called in the literature *Homo erectus* sensu lato (e.g. Antón, 2003:153).

47 This would explain the relatively late entry in Europe as opposed to Asia, given the large European predators of the time, which disappeared only after 1mya (Dennell:2003:423, 431).

48 The climatic "Mid-Pleistocene Revolution", whereby the older glacial-interglacial rhythm of 41ky

Homo populations seemed to have become permanent colonists of the non-African Old World (possibly earlier in South-East Asia) (Dennell, 2003:432).

Probably a variant of this intermittent early occupation model is true, and it is highly reminiscent of a meta-population model[49], and, as opposed to the "classical" (implicit early colonization equals continuous habitation), better explains the coherence of the *Homo erectus* as a species throughout such vast expanses of space and time, allowing it to evolve as a unit while conserving regional features (local adaptations or results of genetic drift), despite presumably low population densities.

By ~1mya, or slightly later, we have definite proof of stable *Homo* populations throughout the Old World (Dennell, 2003; Antón, 2003), which suggests that their cognitive and technological levels were quite impressive by that time. It seems that by ~790kya *Homo erectus* controlled fire in the Levant (Goren-Inbar *et al.*, 2004), that it was an active scavenger and/or hunter (Dennell, 2003:423; Lewin, 1998:351-361), and that they possessed an impressive stone toolkit[50]. Recent work on *Homo erectus* brain and development seems to suggest that it was similar to modern humans: Steven Leigh studied the Mojokerto juvenile *Homo erectus* (Leigh, 2006) who died aged approximately one year (range 0.5-1.5 years, Leigh 2006:104). This specimen has a 663cm^3 endocranial volume (Leigh 2006:104) and falls within the lower 95% regression interval for modern humans and outside the chimp distribution (Leigh 2006:106), suggesting that "*H. erectus* brain growth rates either matched or exceeded those of *H. sapiens* [...] imply[ing] similarities in early life history parameters between *H. sapiens* and *H. erectus*" (p. 107), including a possible adolescent growth spurt and altriciality.

A very interesting piece to the *Homo erectus* puzzle is represented by the discovery of stone tools, dated to 0.8-0.9mya, on the Indonesian island of Flores (Morwood *et al.*, 1998; O'Sullivan *et al.*, 2001; van den Bergh, de Vos & Sondaar, 2001; Morwood *et al.*, 2004). This island belongs to the biogeographical region of Wallacea (Morwood *et al.*, 1998; Storm,

---

changed around 0.8mya to a higher-amplitude 100ky cycle (Wilson, Drury & Chapman, 2000:149-152).

49 A meta-population model envisions a set of ephemeral populations connected by gene flow, with frequent recolonization (Section 2.2.8.2).

50 Oldowan (from ~2.5mya in Africa) and the later, more symmetrical bifaces of Acheulian (starting ~1.6mya also in Africa) (Hurles & Tyler-Smith, 2004:246-248; Lewin, 1998:343-349).

2001; van den Bergh, de Vos & Sondaar, 2001), which is separated from Eurasia by deep waters. The Sunda continental shelf to the north (comprising the Malay peninsula and the Indonesian islands of Sumatra, Java, Bali, Borneo and some other smaller islands) and the Sahul shelf to the south (Australia, New Guinea and Tasmania) each formed single landmasses when sea levels were lowered during glacial maxima (Storm, 2001; van den Bergh, de Vos & Sondaar, 2001; Wilson, Drury & Chapman, 2000). Wallacea, positioned between them, is composed of the islands of Flores, Lombok, Komodo, Sulawesi, Halmahera and others and, during the last 2my, the sea level was never low enough to connect it to either Sunda or Sahul (Storm, 2001:365; Morwood *et al.*, 1998). The geographical configuration of Sunda, Wallacea and Sahul is represented in Figure 2.

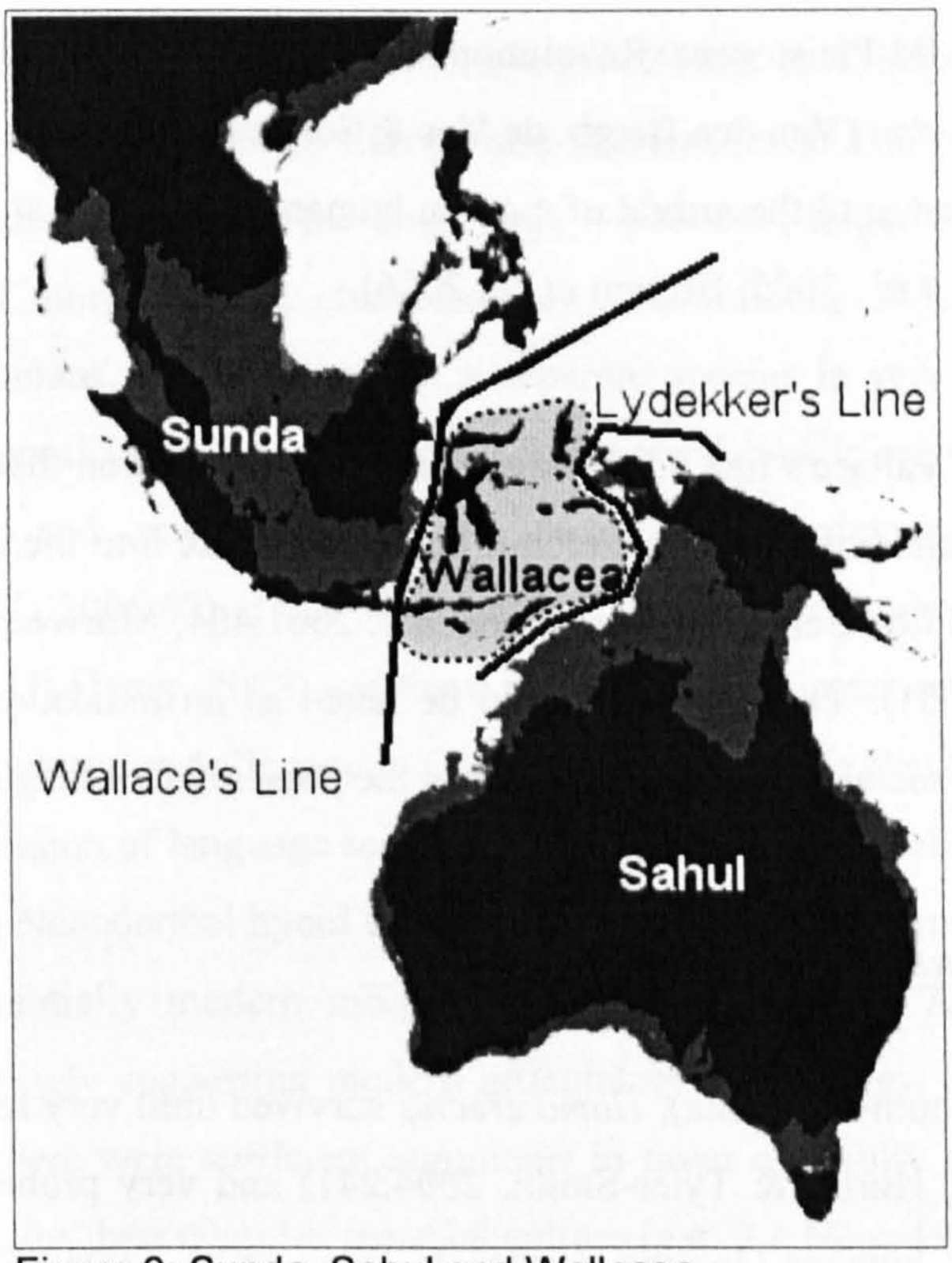

Figure 2: Sunda, Sahul and Wallacea.

Wallace's Line separates Wallacea from Sunda and Lyndekker's Line separates Wallacea from Sahul. During the last 2 my, Wallacea was never connected to Sunda nor Sahul by landbridges.

Thus, the Wallacea's islands could be colonized only by sea-crossings (~50 km during periods of low sea level separating Borneo – belonging to Sunda – and Sulawesi – belonging

to Wallacea - van den Bergh, de Vos & Sondaar, 2001:395). This is also proved by the endemic and non-equilibrated character of these islands' faunas (van den Bergh, de Vos & Sondaar, 2001:404; Strom, 2001:365), because the only species capable of colonizing it were those able to cross water by swimming, rafting or flying (Morwood *et al.*, 1998:174). Concerning specifically Flores, at the lowest sea level, there were still 19 km of water to be crossed (Morwood *et al.*, 1998:176), with very strong surface currents (van den Bergh, de Vos & Sondaar, 2001:404), which made it quite hard to reach for land vertebrates. Thus, the presence of endemic pygmy elephant (*Stegodon*), giant rats and reptiles (Morwood *et al.*, 1998:176), suggests a prolonged isolation of this island. Van den Bergh, de Vos & Sondaar (2001) suggest that "the arrival of the first humans on the island of Flores around 0.8 Ma, coincides with a marked faunal turnover, not only regionally but also worldwide" (p. 404), connected to the "Mid-Pleistocene Revolution" (Footnote 48). It seems that these early humans hunted *Stegodon* (Van den Bergh, de Vos & Sondaar, 2001:405). The island appears continuously inhabited until the arrival of modern humans (Morwood *et al.*, 2004; Brown *et al.*, 2004; Morwood *et al.*, 2005; Brumm *et al.*, 2006).

The crossing of the Wallace's line 0.9-0.8mya is a strong suggestion that *Homo erectus* was capable of building and controlling watercraft able to navigate into the open sea (Morwood *et al.*, 1998:176; van den Bergh, de Vos & Sondaar, 2001:404; Morwood *et al.*, 2004:1091; O'Sullivan *et al.*, 2001). This, in turn, could be taken as an indication of this hominin's cognitive capacities, social organization and, even the possession of language.

2.1.2.3. Homo neanderthalensis

In Asia (especially South-East Asia), *Homo erectus* survived until very recently (e.g., in Java until 27kya; Jobling, Hurles & Tyler-Smith, 2004:241) and very probably interacted with anatomically modern humans (Jobling, Hurles & Tyler-Smith, 2004:241). Throughout this enormous timespan, it seemingly kept evolving, such that the latest specimens tend to fall into the upper range of cranial capacity (1200cm^3), even if the species-specific morphology persists barely altered (Antón, 2003:135, 144).

The later hominins from Africa and Europe are ascribed to *Homo heidelbergensis*, also widespread and variable, probably derived ~1mya from the African *erectus* stock (Jobling,

Hurles & Tyler-Smith, 2004:241; Antón, 2003). They have larger brains, and generally show more *sapiens*-like features[51]. In Europe, there is still controversy concerning the status of *Homo antecessor* discovered at Gran Dolina, Atapuerca (northern Spain), but it seems probable that it should be understood as a variety of *Homo heidelbergensis* (Aguirre & Carbonell, 2001; Jobling, Hurles & Tyler-Smith, 2004:241; Finlayson, 2005). Its contribution to the later *Homo neanderthalensis* of Eurasia is disputed, but probably limited (Aguirre & Carbonell, 2001:14; Finlayson, 2005:458).

The Neanderthal[52], or *Homo neanderthalensis*, was an Euroasian hominin, attested from ~250kya or earlier to as recently as 30-28kya in the Iberian Peninsula (Jobling, Hurles & Tyler-Smith, 2004:241; Zilhão & Trinkaus, 2003b). His range expanded and contracted, following the climate fluctuations of the Pleistocene (Zilhão & Trinkaus, 2003b; Stewart, 2005), but remained circumscribed to Europe and Western Asia. The anatomy was robust, cold adapted[53] (Trinkaus, 1981) and the brain large (~1400cm^3, larger than *Homo sapiens*'). There is a series of morphological characteristics differentiating *Homo neanderthalensis* from the modern humans, but his status as a separate species is very much debated. His cognitive and behavioral capabilities are also controversial, but it seems at least that they were skillful hunters and scavenging was not an important component of their behavior (e.g., Bocherens *et al.*, 2005:83). Their stone toolkit was highly complex (Jobling, Hurles & Tyler-Smith, 2004:247; Henry, 2003) and there are a number of intentional burial sites, some probably also containing symbolic goods (Valladas *et al.*, 1987; Pettitt, 2002). The issue concerning the possession of language seems to have been somehow settled by the discovery of a (very probably) Neanderthal hyoid bone in the Kebara Cave in Israel (Arensburg *et al.*, 1989), showing essentially modern morphology (Fitch, 2000:262; Arensburg & Tillier, 1991) and, thus, strongly suggesting modern articulatory capabilities. But even before this seminal discovery, there were sufficient arguments in favor of a fully articulated language, necessary to account for their complex material culture (e.g., Le May, 1975).

But, partially due to the historical accident represented by the first reconstruction of an old, arthritic, Neanderthal (Jobling, Hurles & Tyler-Smith, 2004:243) and the natural inclination

---

51 In the palaeoanthroplogical literature this is usually referred to as "advanced", still betraying a man-centered *Weltanschauung*. "Advanced" and "sapiens-like" will be used interchangeably, understanding that they are simply descriptive and do not imply any directionality (teleology).

52 See the note on spelling in Relethford, 2003:75 – I will use the old *Neanderthal* form.

53 But see Stewart (2005), which argues this is an adaptation to closed environments.

of modern humans to see themselves as the pinnacle of the living, *Homo neanderthalensis* continues to be depicted as a brutish, decayed, not-quite-human shadow of us, a lost soul on the triumphant march towards humanity, which we alone were able to fully reach (Relethford, 2003). Of course, given what we now know about this extinct hominin, all this is pure nonsense. For example, Stephen Oppenheimer writes:

> [s]o, it could just be that, as some people claim, the beetle-browed appearance of some rugby internationals and soccer hooligans eventually turns out to be a Neanderthal throwback, rather than the more likely event (in my view) of normal variation in modern humans (Oppenheimer, 2004: 49),

while in the Science-Fiction novel *Evolution* by Stephen Baxter (2003), otherwise admirably written, the Neanderthals are depicted as subhuman animals, domestic slaves of modern humans. A more subtle dismissal of the Neanderthals can be found for example in Stringer & McKay (1996:93), whereby they are relegated to an off-shot, a dead-end of human evolution which also *devolved* their arguably most human feature, namely their capacity for language[54]:

> The reasons for Neanderthals' apparent vocal backsliding may be quite straightforward [, specifically reducing the volume of the vocal tract so that] smaller mouthfuls of that freezing European atmosphere [would have been taken, protecting thus the] throats' and lungs' delicate membranes (Stringer & McKay, 1996:93)[55].

The other extreme is represented by such popular novels as John Darnton's *Neanderthal* (Darnton, 1996), where *they* have tremendous para-psychic powers, including mind-reading and the like. Of course, as they are so close to use and still (probably) different, it is not hard to understand the popular fascination with this "alternative" humanity.

After almost 200ky of successful survival and adaptation, *Homo neanderthalensis* disappeared in a matter of some 20ky (Mellars, 2005; Tattersall & Schwartz, 1999; Trinkaus & Zilhão, 2003), process which generated even more interest, both scientific and popular, than their very existence. The arguably scientific hypotheses range from total replacement by incoming modern humans (the standard ROA model, Stringer & Andrews, 1988) by various means [direct or indirect competition for resources (Horan, Bulte & Shogren, 2005; Hockett & Haws, 2005), differential resistance to disease, slight demographic advantage (Hockett & Haws, 2005), sheer bio-cultural superiority, usually language (Stringer & McKie, 1996:93-94)] to absorption into the larger modern human gene pool (Trinkaus & Zilhão, 2004), to

54 Represented by a more elevated position of the larynx than in the last common ancestor of *Homo neanderthalensis* and *Homo sapiens*, postulated on circumstantial evidence.

55 I must note, however, that this represents one of the worst evolutionary explanations to date.

climatically mediated extinction (Stewart, 2005).

My own view is that *Homo neanderthalensis*, far from being a dead branch of the hominin tree, "a 'derived' species, a specialist, one-off diversion from the main hominin line that evolved from *Homo erectus* to *Homo sapiens*" (Stringer & McKie, 1996:93), was, in fact, just a regional variant (Jolly, 2001:193-194) of archaic *Homo sapiens*, belonging to the same biological species and virtually as human as we are, except probably for some slight differences of degree and not of quality. From an anatomical point of view, their external appearance would have been absolutely human and

> [...] in those features of the Neanderthals that would have been accessible for observation – stature, skin, eye and hair color, shape of the face and forehead - [...] would not have fallen outside the casually perceived range of variation [...] among early modern humans (Trinkaus & Zilhão, 2004:548),

while, behaviorally and cognitively, they would have at least matched the moderns. Therefore, building on the currently available evidence[56], I submit that *Homo neanderthalensis* was a regional version of modernity, absorbed and swamped by the incoming African populations. It is highly probable that they contributed genes into the modern gene pool, possibly connected to local adaptations to the temperate or peri-glacial climate of Europe and Western Asia[57].

2.1.3. The evolution of modern humans: the competing models

Historically, there are three main classes of models explaining the emergence of modern humans. Of these three[58], Carleton Coon's *polygenism* is definitively disproved, while the surviving two, Milford Wolpoff's *multiregionalism* and Chris Stringer's *monogenism* are undergoing dramatic evolutions in the light of current discoveries, seemingly converging slowly towards a common model[59].

Coon's **polygenism** (also known as **the candelabra model**) can hardly be considered a

56 See also the discussion of the genetic data (Section 2.2.7).
57 See arguments for the continuity of some Neanderthal features into early modern populations (Section 2.2.5).
58 I will use the names of the most prominent figures supporting the prototypes of these classes of models, even if it can be argued that a long list of prefigurations and alternatives exist (see Wolpoff & Caspari, 1997).
59 This gradual change is detectable when comparing their early (mid to late 80s), recent (90s) and current versions.

scientific theory of human evolution by modern standards (Jackson, 2001; Lewin, 1998; Koller, 2005; Wolpoff & Caspari, 1997) and has been definitively disproved[60]. In 1962, he published *The Origin of Races* (Coon, 1962), where he amassed, with impressive erudition (Stringer & McKie, 1996:46; Dobzhansky, 1963:360), arguments in favor of his theory that, after the migration of *Homo erectus* out of Africa, these populations evolved *independently* towards modern forms "not once but five times[61], as each subspecies, living in its own territory, passed a critical threshold from a more brutal to a more sapient state" (Coon, 1962:658). Moreover, these races attained the *sapiens* status at different times, broadly the Caucasoids and Mongoloids got there first, ~250kya (Stringer & McKie, 1996:46), while the Africans reached this apex only yesterday by evolutionary standards and "the Australian aborigines are still in the act of sloughing off some of the genetic traits which distinguish *Homo erectus* from *Homo sapiens*" (Coon, 1962, cited in Stringer & McKie, 1996:46). As a very well-known example of unfairness and subjectivity, the book pictures (Plate 32) side by side, an Australian aborigine woman and a Chinese academic, while the caption runs:

> The Alpha and Omega of *Homo Sapiens*: An Australian aboriginal woman with a cranial capacity of under 1,000 cc. (Topsy, a Tiwi); and a Chinese sage with a brain nearly twice that size (Dr. Li Chi, the renowned archaeologist and director of the Academia Sinica).

Coon's theory can be represented graphically as in Figure 3 below, which is based on Coon (1962) and comments in Stringer & McKie (1996), Jackson (2001), Lewin (1998) and Koller (2005)[62]. It should be noted that since Coon's time, the history of our genus has been amply revised. I have also chosen to represent the migration out of Africa of *Homo erectus* as roughly simultaneous in all directions. The vertical lines represent Coon's "races" evolving through time in each of the five locations considered by him. I tried to depict his teleological evolutionary process by representing in black[63] "full" *Homo erectus* and in white "full"

---

60 I still discuss it because of its historical importance in shaping the ensuing human evolutionary debates and because it represents a class of theories which *must not* be rediscovered, being already falsified.

61 The five races were the "Caucasoid, Mongoloid, Australoid, Congoid, and Capoid" (Coon, 1962:3), roughly corresponding to Europe, Asia, Australia (plus PNG) and Africa (distinguishing San).

62 The marked times/dates, migratory routes and rates of change from *erectus* to *sapiens* are chosen for illustrative purposes only.

63 There is no hidden racist nuance in this graphical representation, as some decided readers might be inclined to detect. It is solely imposed by the requirement that the drawing must use shades of gray only and the fact that if white would have been used to represent *Homo erectus* and black *Homo sapiens*, then the horizontal "migratory" lines into the Old World would have been also white and vary hard to see.

Homo sapiens, while the various shades of gray represent intermediary stages on Coon's transformation of *erectus* into *sapiens*. Conforming to this theory, ~1mya, there was a single species, *Homo erectus*, living in Africa[64], which, "over half a million years ago" (Coon, 1962:657) spread around the Old World[65] and started differentiating into his five races at different times and at different rates[66]. The first to have begun this process are, conforming to Coon, the Caucasoids (Europe) and Mongoloids (East Asia), and they reached full *sapiens* status ~250kya[67], followed more than 200ky later (Jackson, 2001:248) by the Congoids and Capoids (Africa), while the Australoids (Australia and PNG) haven't yet fully attained this stage[68]. The figure also emphasizes Coon's insistence that there was no contact between these evolving lineages and thus, the entire drawing reassembles an upside down candelabra.

This theory was immediately criticized (e.g. Dobzhansky, 1963; Montagu, 1963) on various grounds, including the fact that *Homo sapiens* is overwhelmingly uniform, that the teleological, directed independent evolution towards modernity is untenable on biological grounds, that his assumptions about brain size are simplistic[69] and that his personal political views influenced his scientific judgment (Dobzhansky, 1963; Montagu, 1963; Stringer & McKay, 1996; Lewin, 1998; Wolpoff & Caspari, 1997). Unfortunately, this view of human evolution, though patently wrong even by the time it was published, fueled the extreme conservative movement, by apparently justifying a segregationist politics. Also, ironically, and unfortunately for science, it provoked a repulsive reaction so strong that the views changed into the opposite extreme: there are *no differences* whatsoever, modern human populations are *uniform* in all relevant aspects, and the attempt to study variation is both *futile* and *intrinsically dangerous* (Wolpoff & Caspari, 1997). The ensuing history of palaeoanthropology, I believe, can be understood (partially, at least) as a running away as far as possible from Coonian ideas.

---

64 Represented by the single black vertical lineage starting from the bottommost plane.

65 The black winding lines in the middle plane, spreading from eastern Africa towards southern Africa, east Asia, Europe and Australia.

66 This is represented in the figure by the grading shades of gray from black towards white in the five vertical lines starting from the middle plane.

67 Represented in the figure by their pure white vertical lines, starting 250kya.

68 Represented in the figure by the still gray (non-white) gradient of their line.

69 See for example, the dispute between Coon and Montagu concerning Anatole France's small brain, comparable to Topsy's (Montagu, 1963:364).

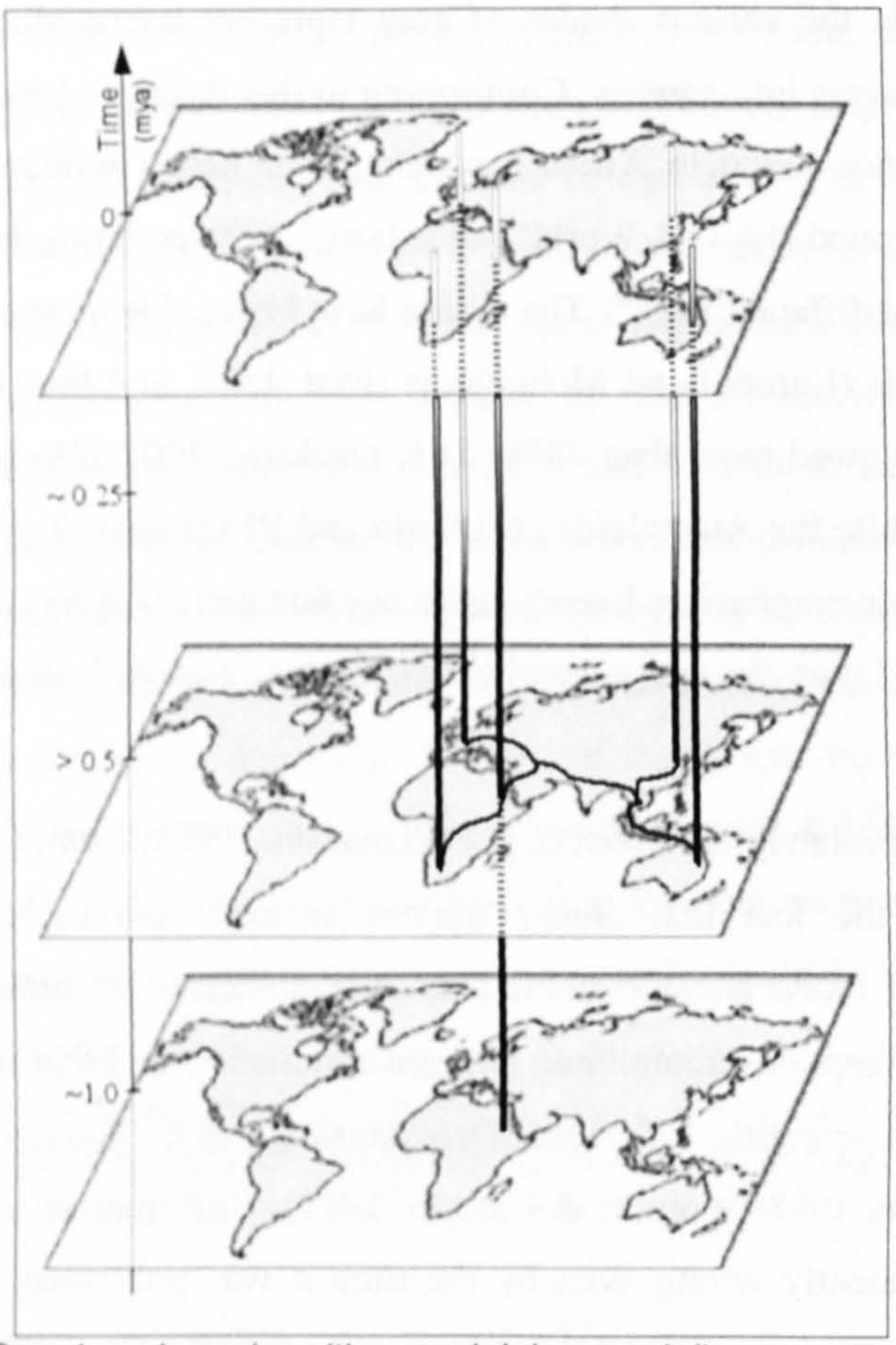

Figure 3: Carleton Coon's polygenism (the candelabra model).
This drawing is based on Coon (1962) and comments in Stringer & McKie (1996), Jackson (2001), Lewin (1998) and Koller (2005). The vertical axis represents time in million years before present (mya). See text for details and explanations.

Milford Wolpoff's **multiregionalism** is very difficult to summarize, both because it is an evolving entity, and because it was portrayed in so many different ways [most of them, only tenuously connected to its core ideas, and some of them, simply wrong (Wolpoff, Hawks & Caspari, 2000; Eckhardt, Wolpoff & Thorne, 1993; Relethford, 2001:63-65; Wolpoff & Caspari, 1997)]. The picture is also complicated by the widespread idea that it had been falsified, representing a dead model, a thing of the past, and its supporters an "isolated, albeit vociferous minority" (Oppenheimer, 2004:50). I will present my own understanding of the "classical" form of this model, mostly based on Milford Wolpoff and Rachel Caspari's book *Race and Human Evolution: A Fatal Attraction* (Wolpoff & Caspari, 1997)[70], as well as

70 Its first edition (1996) won the 1997 American Anthropological Association's (Biological Anthropology Section) W. W. Howells Book Prize (http://www.as.ua.edu/bas/BookPrize.htm September 2006).

Thorne & Wolpoff (2003), Relethford, (2001), Eckhardt, Wolpoff & Thorne (1993), Wolpoff, Hawks & Caspari (2000), Wolpoff & Caspari (2000) and Hawks & Wolpoff (2001, 2003).

The theory was first articulated in relation to the fossil record of East Asia (Wolpoff, Wu & Thorne, 1984) and was originally designed to explain the puzzle of regional features continuity over very long time spans, coupled with a constant trend towards modernity. It was based on Franz Weidenreich's[71] "*polycentric* theory of human origins", formulated as early as 1938 (Eckhardt, Wolpoff & Thorne, 1993:974). Weidenreich's views are usually misunderstood as being polygenic and pooled with Coon's (for example, Stringer & McKie, 1996:46, 48; Cavalli-Sforza, Menozzi & Piazza, 1993:639; Cavalli-Sforza, Menozzi & Piazza, 1994:63; Hanihara, 1996:389, 391), while, in reality, Coon was Weidenreich's student, but did not continue his scientific ideas[72]. Maybe there is no better illustration that the two men had very different views, than that the geneticist Theodosius Dobzhansky, one of the most fervent critics of Coon's polygenism, was strongly influenced by Weidenreich's ideas of regional continuity and global contact[73] (Dobzhansky, 1944; Hawks & Wolpoff, 2003:89; Tattersall, 2000:3). This point is clearly made by Roger Lewin:

> Weidenreich was aware that [...] modern races might be considered to have separate origins, even to be separate species. In 1949 [...], he explicitly ruled out this possibility. In fact, in 1962 [...] Carleton Coon came close to proposing the hypothesis against which Weidenreich had warned (Lewin, 1998: 378-379).

To disperse this confusion, which lasted too long both inside and outside palaeoanthropology, I think it is best to start by making clear what multiregionalism is *not*:

- *it is not a theory of multiple origins*: it posits that after the *Homo erectus* expansion out of Africa, the lineage evolved as a unitary entity, the regions being connected through gene flow. Thus, there is a single deep origin of modern humans in Africa, but asking about the recent origins of modern humans is in a way meaningless, as the answer is *everywhere* and *nowhere* specifically;
- *it does not involve parallel evolution*: because there has been constant gene flow between populations, features could spread throughout the human species, insuring concerted

71 Weidenreich's biographies are in Wolpoff & Caspari (1997), Gregory (1949) and Haviland (2000:238).
72 Even if his book *The Origin of Races* was dedicated to Weidenreich (Stringer & McKie, 1996:46).
73 Extremely telling is the fragment cited in Hawks & Wolpoff (2003:89) from Dobzhansky (1955:3-4).

evolutionary change;

- *it is not about Neanderthals being the ancestors of modern Europeans*: it is possible that Neanderthals contributed genes into the Upper Paleolithic European gene pool but this does not make them *the* ancestors. Thus, the debate concerning their actual contribution to the modern gene pool does not bear on the theory as a whole;
- *it does not assume equal contribution by all geographic regions into the modern gene pool*: it is to be expected, that by sheer population size and time depth, African contribution was overwhelmingly important compared to Europe or East Asia.

Multiregionalism is also called a *trellis* model of human evolution, because of the horizontal gene flow links between the main geographic regions. It is visually represented in Figure 4. I have considered the same five regions as in Coon's polygenism depiction, in order to contrast the models better. The intersecting lines, connecting these regions throughout the evolutionary history of *Homo* represent gene flow. To simplify the drawing, I have pictured only flow between neighboring regions. Also, the thicker connections do not represent the amount of gene flow but only a visual aid to distinguish those near the viewer from those farther away. Another simplification concerns the dates of *Homo erectus/sapiens* migrations to various parts of the Old World, which are represented as simultaneous and as initiating permanent settlement. Noteworthy in this respect is Australia/PNG, where *Homo sapiens* arrived only after ~60kya. The most obvious and important consequence of this model is that it is meaningless to talk about the place of origin of modern humans or of various human populations. Moreover, the entire human species evolved in synchrony[74], with advantageous alleles spreading throughout its range. Also, the entire genus *Homo* is composed by only a single biological species, which still can be divided into regional morphs and chronospecies.

Contrary to the apparent consensus, multiregionalism has *not* been invalidated, either by the fossil record, or the genetic data. In fact, even the most fervent proponents of ROA, including Chris Stringer, have begun to admit various degrees of admixture between the expanding waves of modern *Homo sapiens* and the ancestral stocks (see below), which is a covert form of multiregionalism. Moreover, it seems that the apparent success of ROA is due more to a historical accident than to lack of alternative explanations (Sections 2.3 and Annex 2).

74 Thus, all human populations reached modernity at the same time, to use Coonian language.

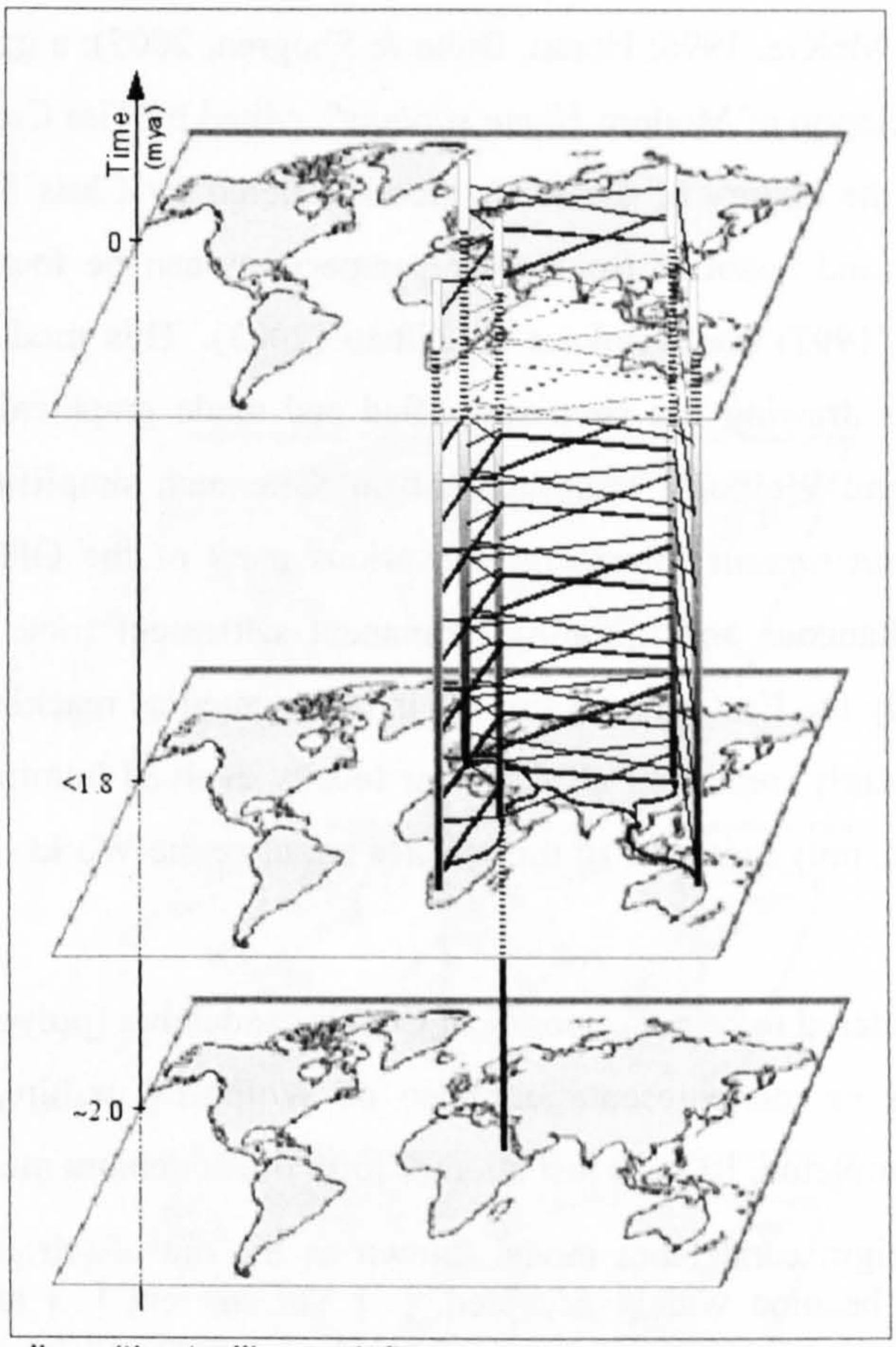

Figure 4: Multiregionalism (the trellis model).
The vertical axis represents time in million years before present (mya). The lines connecting the main geographical populations represent gene flow (the thicker lines do not represent higher amounts of gene flow but simply visual closeness to the viewer). To simplify the drawing, only flow between neighboring regions is represented. See text for details and explanations. Another simplification concerns the dates of Homo erectus/sapiens migrations to various parts of the Old World, which are represented as simultaneous and initiating permanent settlement (note in this respect the Australian case).

Recent out of Africa, or **the replacement model**, primarily due to Chris Stringer and Peter Andrews (Stringer & Andrews, 1988), is generally accepted as the true story of modern humans origins. Its main idea is that *Homo sapiens* appeared recently in Africa, from a local hominin stock, and represented a new biological species, different from the other contemporaneous hominins (including Asian *Homo erectus* and Eurasian Neanderthals), and, later expanded throughout the world, replacing the local archaics because of superior technology and cognition. The speciation of *Homo sapiens* is regarded as an event, happening in one place and at a given time, and hypotheses vary widely but usually focus on a *single* determinant, be it language, large-scale trade, social division of work, etc. (Crow,

2002a, b; Stringer & McKie, 1996; Horan, Bulte & Shogren, 2005): a good such example is offered by "The Speciation of Modern *Homo sapiens*", edited by Tim Crow (Crow, 2002). A personal account of the history of the ROA idea is offered by Chris Stringer (Stringer & McKie, 1996:65-83) and histories from other perspectives can be found, for example, in Wolpoff & Caspari (1997) and Trinkaus & Zilhão (2003). This model is represented in Figure 5 below. This drawing has been simplified and made graphically compatible with Coon's polygenism and Wolpoff's multiregionalism. One such simplification concerns the dates of *Homo erectus/sapiens* migrations to various parts of the Old World, which are represented as simultaneous and initiating permanent settlement (note in this respect the Australian case). Only the East African lineage is represented as reaching the *sapiens* form (white) and, subsequently, replacing all the other locally evolved hominin lineages (black), so that, after ~150kya, only members of this species populate the World.

ROA is usually considered to be the opposite of Coon's candelabra (polygenism) model (and, by misunderstanding or misrepresentation, also of Wolpoff's multiregionalism), but, as observed by Alan Templeton, ROA is just another form of candelabra model:

> [...] a recent origin candelabra model known as the out-of-Africa replacement hypothesis has become widely accepted. [...] The ancient [...] and recent [...] candelabra models differ only in their temporal placement of the ancestral node but share the same tree topology that portrays Africans, Europeans and Asians as distinct branches of an evolutionary tree (Templeton, 1998:636).

This observation is valid, and usually implicit in discussions of human evolution, by depicting the modern populations after their split as branches of an evolutionary tree (Templeton, 1998; Templeton, 2002; Hawks & Wolpoff, 2001:44). Given that one of the main thrusts of ROA in the media was its supposed anti-racist implications, as opposed to the perceived racist and conservatory multiregionalism (Wolpoff & Caspari, 1997; Annex 2), the interpretation of ROA as a candelabra model weakens this argument (Templeton, 1998; Wolpoff & Caspari, 2000; Wolpoff & Caspari, 1997). As extensively discussed (Templeton, 1998; Wolpoff & Caspari, 2000; Wolpoff and Caspari, 1997; Banton, 1998), human races are *not* valid representations of human diversity not merely because of their recency (as ROA posits) but because *they do not represent independent evolutionary lineages*, to which the question of age is meaningless (as multiregionalism strongly asserts). Thus, the accusations of racism towards the trellis model are a consequence of misunderstanding and misrepresentations of both ROA and multiregionalism.

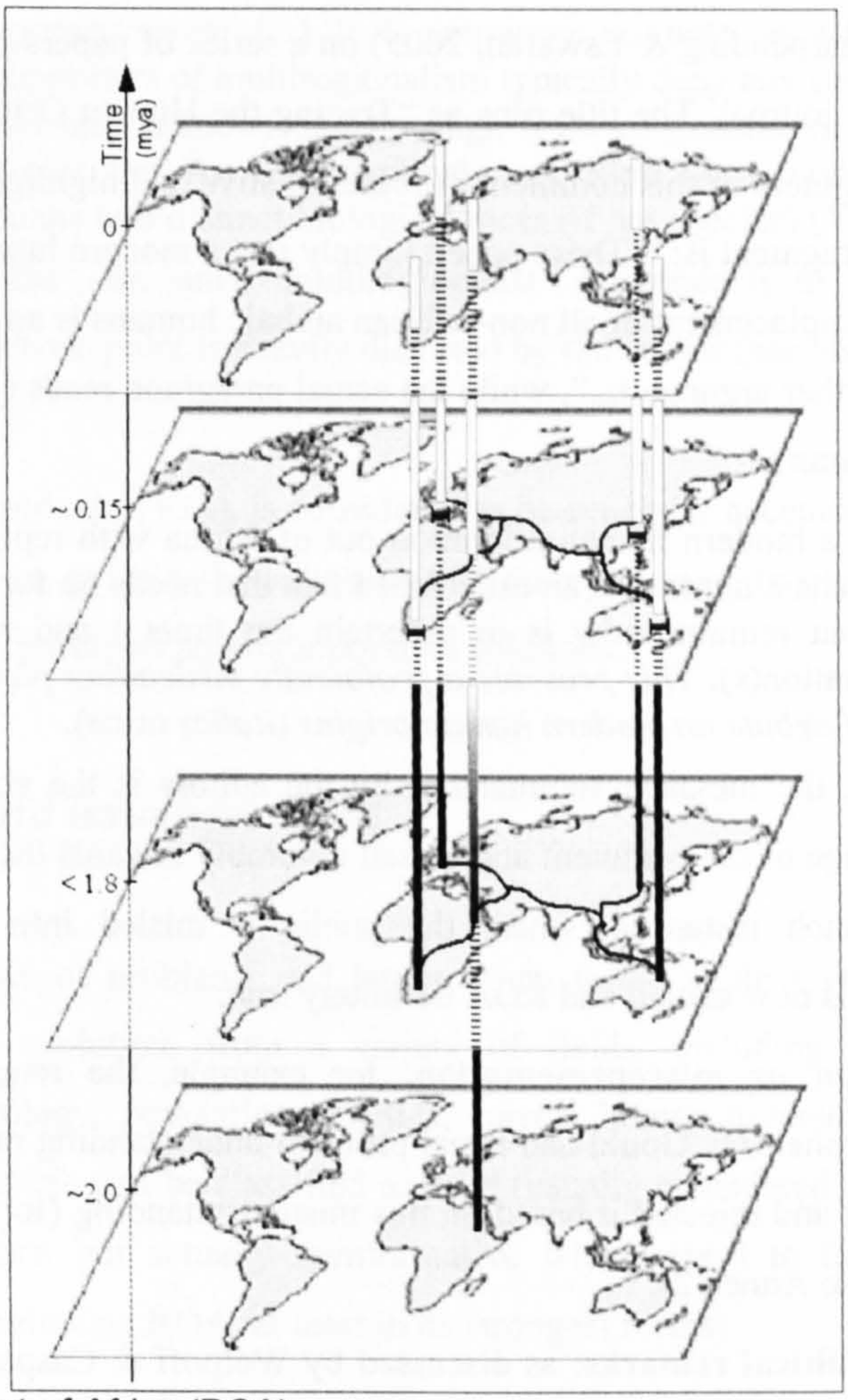

Figure 5: Recent-Out-of-Africa (ROA).

The vertical axis represents time in million years before present (mya). Homo sapiens emerges from the lineage evolving in East Africa and replaces all the other hominin populations (represented by black vertical lines interrupted by horizontal cuts and replaced by white vertical lines). This drawing has been simplified and made graphically comparable to Coon's polygenism. One such simplification concerns the dates of Homo erectus/sapiens migrations to various parts of the Old World, which are represented as simultaneous and initiating permanent settlement (note in this respect the Australian case).

During the last decades, the controversy concerning modern human origins has been bitter, in many ways even non-academic. Another complicating factor in assessing these alternative models is represented by what looks like a marketing program directed towards the public and containing misrepresentations, caricatural simplifications and political or moral assertions. I will present just a small sample below:

- **biased representation** in the popularization and scientific press: for example, on the page 1995 of the 23rd of September, 2005 number of *Science* (volume 309), there is a

comment (Harpending & Eswaran, 2005) on a series of papers previously published by the same journal. The title runs as "Tracing the Human Origins" and the editors picked a fragment of this comment as representative and highlighted it[75] on the same page. This fragment is: "[These papers] imply that a modern human migration out of Africa with replacement of all non-African archaic humans is an established fact that needs no further argument...", while the actual paragraph reads (page 1995, lines 10-18, first column):

> [...] imply that a modern human migration out of Africa with replacement of all non-African archaic humans is an established fact that needs no further argument, and that all that remains now is to ascertain the time(s) and route(s) of the purported migration(s). *This presents a profoundly misleading picture about the present state of debate on modern human origins* (*italics* mine).

In this case, the message summarized by the editors is the sheer opposite of the actual message of the comment and biased favorably towards the ROA model. There are many such instances, where the public is misled into believing that the controversy is now settled and ROA definitely true;

- **simplification or misrepresentation:** for example, the renowned evolutionary biologist Stephen Jay Gould had a real problem understanding multiregionalism as a trellis model and rejected it based on this misunderstanding (for example, in Gould, 2002; see also Annex 2);
- **moral or political remarks:** as discussed by Wolpoff & Caspari (1997; 2000) and Templeton (1998), multiregionalism is sometimes presented as a politically biased, conservative justificatory theory, supporting racial discrimination and white superiority. Possibly one of the best (or worst) such cases is Chris Stringer and Robin McKie's *African Exodus* (1996), where they actively confuse multiregionalism and Coon's polygenism, and accuse the first that "[s]uch a theory would suggest, at face value, that modern humanity's constituent races are divided by fundamental and deep-rooted differences." (Stringer & McKie, 1996:49; but see pages 48-50 for the entire discussion).

A very transparent example of indirect rejection of multiregionalism based on political assertions (disguised as scientific opinions but lacking any real basis), is represented by Richard York's paper (York, 2005), where he claims that:

> Multiregionalists adhere to the position that the division of humans into distinct groups (races) is very old, which implies that genuine biological differences exist

75 Blue, large font.

> among contemporary races. [...] It is important to note in all fairness that contemporary supporters of multiregionalism typically deny any support for racist views or policies and acknowledge the high level of genetic similarity among human populations, but the multiregionalist position does, nonetheless, reify divisions of humans into distinct biological races (if not species) (York, 2005).

"In all fairness", the only "multiregionalist" reference is to Wolpoff & Caspari (1997), but whose point is heavily distorted by the author (see also Annex 2).

But, no matter how and why, ROA is considered to be generally accepted, especially outside palaeoanthropology, or, at least, the only scientific theory at hand. In the following section, I will analyze its main problems and their possible solutions.

2.2. Problems and issues for ROA

ROA reveals a series of problems and issues when tested against various types of data relevant to human evolution, from a variety of fields, including palaeoanthropology, archaeology, primatology, population genetics, current human diversity and ancient DNA studies. In general, these can be classified as *mild* (usually considered to support ROA and falsify its competitors, but actually uninformative with respect to this controversy) and *serious* (potentially rejecting ROA, at least in its strongest forms).

2.2.1. The transition to *Homo sapiens* was not a "revolution"

In March 2000, a meeting was dedicated to debating the "topic of the speciation of modern *Homo sapiens*" (Crow, 2002a:1) and resulted in the publication of Crow (2002). The human evolution section was heavily biased (Chris Stringer, Paul Mellars and Ian Tattersall, well-known advocates of ROA and the separate species status for *Homo sapiens*), making the absence of more equilibrated participants (e.g, John Relethford) or opponents (e.g., Milford Wolpoff or Sally McBrearty) even more obvious. Moreover, even from the title it was assumed not only that *Homo sapiens* was a biological species, but even its "modern" form[76] was one. In the introduction, Tim Crow states:

> [t]he paradigm of *H. sapiens* therefore suggests a new version of saltational speciation, that it is not chromosomal changes in general that play a role in

76 Whatever that means, if it has any meaning at all (e.g. Stringer, 2002:575).

> speciation but changes on the sex chromosomes, and perhaps changes in regions of X-Y homology that are involved (Crow, 2002a:13)

setting the goal for the entire book, which is trying very hard to find "a *single* gene [that] played a critical role in the transition from a precursor species [to modern *Homo sapiens*]" (Crow, 2002b: 198, *italics* mine), and which, he indeed manages to identify as being the *protocadherinXY* gene located on the X-Y homologous region (Xq21.3/Yp11.2; Crow, 2002b:197-210). But *why* so much trouble for such a hard-to-believe story?

"The search for revolutions in western thought has been in part [...] a search for the soul, for the inventive spark that distinguishes humans from the rest of the animal kingdom" (McBrearty & Brooks, 2000:533). People seem to need clear boundaries separating *them* from *the others*, be it different humans (racism), different social classes, or even the other sex, but especially from the "inferior" hordes of speechless creatures, purportedly created to serve man. In the age of evolution, when it is clear that we do obey the same principles as all the other living things, this boundary has to take the form of a "revolution", a sudden, total, profound change which made us entirely human in just one move. Gradualism won't do, as it allows for intermediate shades of humanity, but a single mutation which gives language, cognition, social structure and everything else will be just perfect. "By stressing human uniqueness, proponents of the "human revolution" effectively remove the origin of *H. sapiens* from the realm of normal scientific inquiry." (McBrearty & Brooks, 2000:533).

Given the Eurocentric view of human evolution, which persisted for various reasons for a long time (McBrearty & Brooks, 2000; Mellars, 2005; Stringer, 2002; Henshilwood & Marean, 2003; Haviland, 2000), there seemed to be some justification for such a "human revolution". The archaeological and fossil record of Europe does indeed show what looks like a rapid succession of two different types of hominins and cultures. The transition appears catastrophic (taking several thousands of years) and important, bringing art, advanced technology, personal ornaments and long-distance trade (McBrearty & Brooks, 2000; Mellars, 2005; Stringer, 2002; Henshilwood & Marean, 2003; Haviland, 2000). Before this pattern could have been appreciated in the larger, world-wide context, it appeared to suggest a saltational event, a sudden "mutation" which produced the full "modern" package, both morphologically and behaviorally. Many proposals have been made to explain this shift, including external memory, conceptual spaces, integration of cognitive modules, contextual focus, economic networks and, most of all, language (Donald, 1999; Klein, 1999;

Mithen, 1996; Gabora, 2003; Dunbar, 1996; Bickerton, 2002; Horan, Bulte & Shogren, 2005).

The trouble is that, as soon as Europe, this "remote *cul de sac*" (McBrearty & Brooks, 2000:454) for human evolution is defocused, a new pattern emerges. The "human revolution" of about 50-40kya (McBrearty & Brooks, 2000:453; Henshilwood & Marean, 2003:629) turns out to be an illusion, an effect of demography, as the "modern" hominins did not suddenly evolve *in situ*, following a catastrophic mutation of some sort, but instead represent an intrusive population coming from elsewhere (McBrearty & Brooks, 2000:454; Mellars, 2005; Henshilwood & Marean, 2003; Haviland, 2000; Stringer & McKie, 1996). Their source is very probably Africa, *via* the Levant, where the picture is very different, a piecemeal accretion of modernity, both morphologically and behaviorally. As argued at length by Sally McBrearty and Alison Brooks in their seminal paper (McBrearty & Brooks, 2000),

> [t]here was no "human revolution" in Africa [rather] [d]istinct elements of the social, economic, and subsistence bases changed at different rates and appeared at different times and places [supporting the view] that both human anatomy and human behaviour were intermittently transformed from an archaic to a more modern pattern over a period of more than 200,000 years (McBrearty & Brooks, 2000:458).

This conclusion is supported by many other studies, both anatomically and behaviorally (Stringer, 2002; Bocherens *et al.*, 2005; Wolpoff & Caspari, 2000; Hawks & Wolpoff, 2001; Wolpoff *et al.*, 2004; Trinkaus *et al.*, 2003; Grün *et al.*, 2005; Henshilwood & Marean, 2003; White *et al.*, 2003; Lee & Wolpoff, 2003; Finlayson, 2005; Eswaran, 2002; Wu, 2003; Morwood *et al.*, 1998; O'Sullivan *et al.*, 2001; van den Bergh, de Vos & Sondaar, 2001; Morwood *et al.*, 2004; Morwood *et al.*, 2005; Falk *et al.*, 2005; Brown *et al.*, 2004; Morwood *et al.*, 2004; Vanhaeren *et al.*, 2006). For example, Lee & Wolpoff (2003) show that the brain size changes through the hominin lineage can be explained by a single continuous process, "incompatible with an interpretation of punctuated equilibrium during this period" (Lee & Wolpoff, 2003:186)[77], and Chris Stringer concludes that "[...] morphological and behavioral evolution were decoupled, since 'morphological modernity'

77 Punctuated equilibrium (Eldredge & Gould, 1972) is usually misunderstood to be the opposite of gradualism, but it represents a reflection of peripatric speciation (Skelton, 1993:393) in the fossil record (Skelton, 1993:489; Berlocher, 1998:10; West-Eberhard, 2003:617-629; Dawkins, 1986). That sudden and profound speciation events can happen is hotly debated and usually explained through phenotypic plasticity (West-Eberhard, 2003, especially 617-629).

may have evolved before 'behavioral modernity' (Stringer, 2002:575).

Christopher Henshilwood and Curtis Marean fully reject the "behavioral-trait list approach" used to identifying modernity in the archaeological record, arguing that:

> [...] many of the traits have several deficiencies. First, they are empirically derived, leading to circularity, and the empirical grounding has its roots in Europe, particularly western Europe [...]. Second, many of the traits can be linked to resource or labor intensification and environmental pressure and thus have nothing to do with the origin of modern human behavior (Henshilwood & Maren, 2003:631)

and conclude that "[...] modern human behavior did not suddenly emerge at ca. 50,000 years ago and cannot be defined by the simple presence or absence of items on a Eurocentrically derived trait list." (p. 637).

Strictly concerning language, for a long time, modern *Homo sapiens* was defined as possessing it to the exclusion of the others. Fortunately, the discovery of a Neanderthal hyoid bone in the Kebara Cave in Israel (Arensburg *et al.*, 1989), showing essentially modern morphology (Fitch, 2000:262; Arensburg & Tillier, 1991), suggests that this is not the case. Moreover, the discovery on the island of Flores of stone tools dating to approximately 800-900kya (Morwood *et al.*, 1998; O'Sullivan *et al.*, 2001; van den Bergh, de Vos & Sondaar, 2001; Morwood *et al.*, 2004), raises the question of the seafaring and colonizing capacities of *Homo erectus* ~1mya and suggest, with a very high probability, the presence of complex language.

The fact that there was no "human revolution"[78], that "modernity" did not represent a homogeneous package, excludes any theory which assumes a speciation *event* as opposed to a *process*, and weakens any claims that *Homo sapiens* represents a distinct biological species somehow "special". Moreover, this mosaic, accretionary, view of human evolution highlights the possibility that various "modern" features represent in fact integrated systems whose components have different origins.

78 However, there are some authors which, even if they accept that the European "revolution" is an illusion, do transfer the concept to Africa, assuming a "human revolution" associated with the MSA and the appearance of *Homo sapiens* (e.g., Mellars, 2005), but their arguments are unconvincing.

2.2.2. A structured population for the origins of *Homo sapiens*

In a paper published in August 2005 in *Genetics* (Garrigan *et al.*, 2005a), a global sample of 42 X (male) chromosomes was analyzed at the Xp21.1 locus. Two African individuals were identified carrying a lineage of non-coding sequence (17.5kb) which seems to not have been recombining with other lineages for more than 1my (Garrigan *et al.*, 2005a:1853). This strongly suggests that this X chromosome lineage evolved in isolation from the other lineages (Figure 6, adapted from Garrigan *et al.*, 2005a:1850):

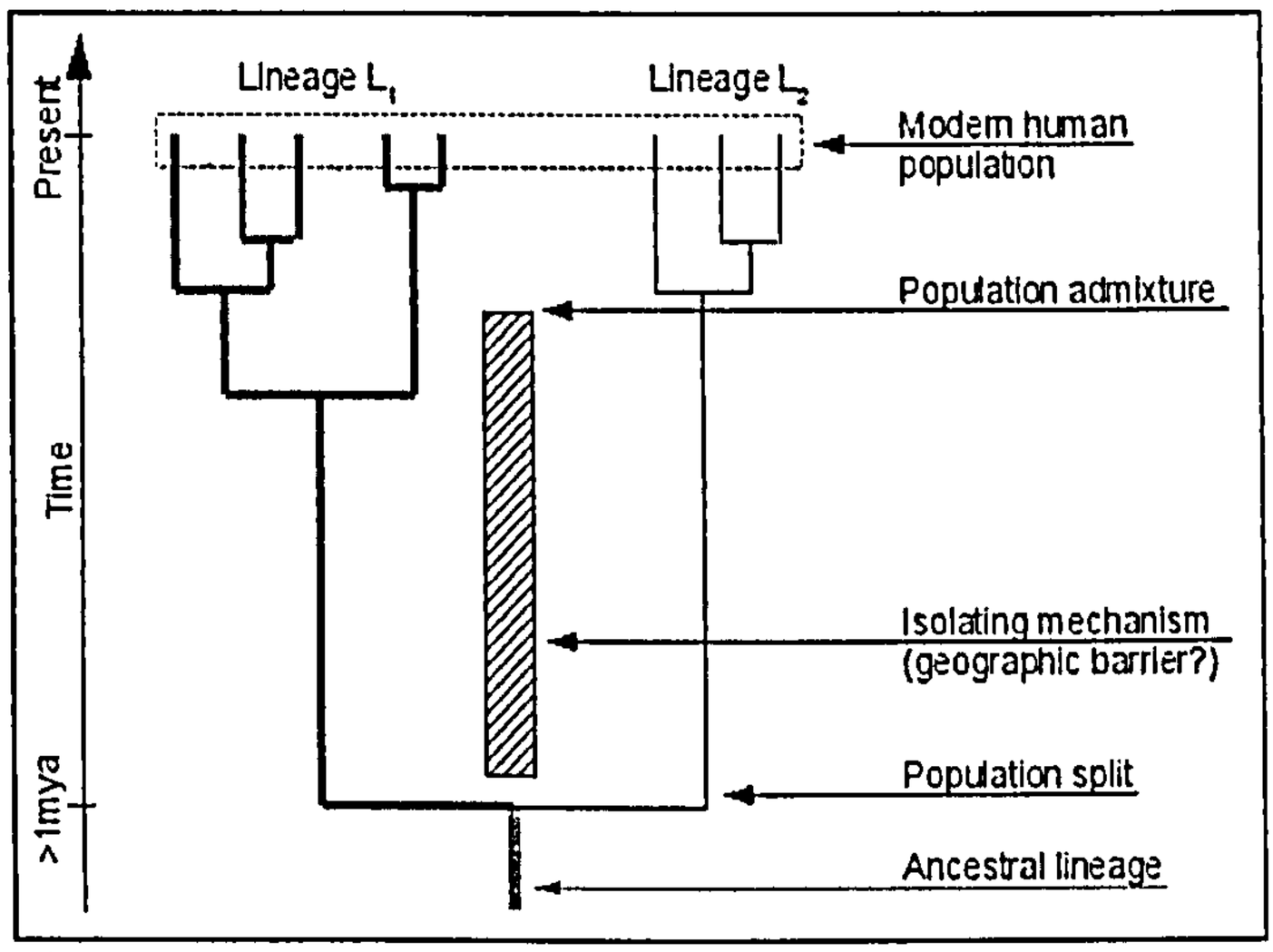

Figure 6: Two lineages of the Xp21.1 locus diverging >1mya and evolving without recombination.

These two lineages (L_1 and L_2) were separated by some isolating mechanism, most probably geographic in nature, for more than 1my and, during this period, they evolved separately, without recombination. At a given moment, these two demes met and admixed, allowing the two lineages to be represented in the modern human population. It is probable that the dissolution of the isolating mechanism was due to a range expansion of one or both of the separated demes, coinciding with the emergence of modern *Homo sapiens* in Africa. It is, thus, assumed that the divergence and subsequent admixture of these lineages took place in Africa and antedates the last major expansion into the Old World (Garrigan *et al.*, 2005a:1855).

This study rejects panmixia with a probability $p < 0.05$ for all meaningful values of the

population mutation and recombination parameters (Garrigan *et al.*, 2005a:1853) and strongly suggests that modern humans originate from admixture of separately evolving lineages. This is a very important find, as it profoundly changes the plausibility of different human evolutionary theories:

> (i)f the AMH [anatomically modern human] genome contains *any* degree of dual ancestry (*i.e.*, archaic and modern), the recent African replacement model in its strictest definition (*i.e.*, that of complete replacement) must be rejected. [...] the evolutionary lineage leading to AMH did not evolve reproductive isolation from other archaic hominin subpopulations and, thus, cannot be considered a distinct biological species (Garrigan *et al.*, 2005a:1855, *italics* in original).

The further possibility that admixture between hominin lineages is not restricted to non-coding genetic material is even more interesting (Garrigan *et al.*, 2005a:1855), but harder to ascertain due to possible selection.

The same theme, that modern humans descend from a structured ancestral population and not a panmictic isolate through allopatric speciation, is reviewed by Harding & McVean (2004) who conclude that the best model accounting for the existing genetic variation is a metapopulational model as opposed to the classical bottleneck/migration with replacement scenario of ROA (Harding & McVean, 2004:671-672; Section 2.2.8). And even Chris Stringer, one of the most important proponents of ROA, concedes: "[...] could there have been an African-based multiregional model where 'modern' behaviours, morphologies and genes coalesced from different parts of that continent during the Middle Pleistocene?" (Stringer, 2002:576). It is important to note, however, that such a concession is not a matter of superficial detail, as usually considered, but a profound rejection of species-status claims for modern *Homo sapiens*, opening the possibility of contributions, both cultural and genetic, of other non-African "archaics" to our modern diversity.

2.2.3. Genes with deep, non-African branches

Most human genes have the deepest branches of their evolutionary trees rooted in Africa, where they also have the highest diversity (Jobling, Hurles & Tyler-Smith, 2004; Relethford, 2001), but the X chromosome seems to behave differently.

The autosomes and the sex chromosomes have different inheritance characteristics, very important for evolutionary studies. The *effective population size*, N_e, is a measure of the

magnitude of random genetic drift, introduced by Sewall Wright (1931), representing the size of an ideal population (random mating, no selection, random chance of each offspring having any particular parent) which experiences the same amount of genetic drift as the population considered (Jobling, Hurles & Tyler-Smith, 2004:131; Halliburton, 2004:236-237; Relethford, 2001:147). It depends on many factors, including population size fluctuations, population substructure and the genetic system under consideration (Jobling, Hurles & Tyler-Smith, 2004:132-133). The long-term N_e is approximated by the harmonic mean of effective population size at different points in time (Jobling, Hurles & Tyler-Smith, 2004:132; Relethford, 2001:148-149), and is extremely sensitive to small values. The relationship between *effective* size, *census* size and *breeding* size for a given population is complex and hotly debated (Relethford, 2001:149). In the ideal case of a Wright-Fisher population, N_e of the Y chromosomes is ¼, and N_e of the X chromosome is ¾ of the N_e for autosomes, and this can reach 1/8 and 9/8, respectively, for populations with extreme variance in male reproductive success (Jobling, Hurles & Tyler-Smith 2004:134, Box 5.1). It is, thus, very important to specify the population model and historical fluctuations for a given genetic system, in order to meaningfully interpret the resulting N_e. The X chromosome could prove very important for human evolutionary studies because of its specific model of inheritance, but, for the moment, it seems to be underexploited (Schaffer, 2004:43).

An ~10kb non-coding region in the locus *HS571B2* on the X chromosome (Xq21.1-21.33) was sequenced in a sample of individuals from Africa, Asia and Europe (Yu, Fu & Li, 2002:2131-2132). A non-African specific variant was found at a frequency of 35% in non-Africans , which could have arisen in Eurasia more than 140kya, predating the appearance of modern *Homo sapiens* (Yu, Fu & Li, 2002:2140-2141). This suggests that "[...]the genetic history at this region [on the X chromosome] in Eurasia may be as deep as that in Africa" (Yu, Fu & Li, 2002:2141) and supports an interpretation of admixture outside Africa between local and expanding African populations.

Another locus on the X chromosome, segments of the *Dystrophin* gene (introns and microsatellites), was sequenced in a 1343 individuals global sample (Ziętkiewicz *et al.*, 2003). One of the three identified lineages consists of a haplotype which is virtually absent from Africa, and seems to be older than the recent expansion (earlier than 160kya). Also, it occupies the position closest to the root in the tree (Ziętkiewicz *et al.*, 2003). This suggests

that the expanding African population admixed outside Africa with local populations, allowing this lineage to survive into the present.

Probably the best such "anomaly" is represented by the *RRM2P4* pseudogene. Garrigan *et al.* (2005b) sampled 41 individuals from a worldwide distribution and found that the reconstructed tree is rooted in East Asia, it has a very ancient MRCA (~2mya) and it also yields a greater non-African than African nucleotide diversity (Garrigan *et al.*, 2005b:190-191) (Figure 7). "The distribution of the Asian lineage strongly suggests an Asian origin but should not be taken as definitive proof that it did not originate in Africa" (Garrigan *et al.*, 2004:191), but it does support a model whereby incoming modern African population admixed with local populations. The divergence time (~2mya) is compatible with the expansion of *Homo erectus*, making it plausible to suggest that the admixture occurred between modern *Homo sapiens* and local *Homo erectus*. This "[...] would have important implications for our view of *Homo sapiens* as a species" (Garrigan *et al.*, 2004:191).

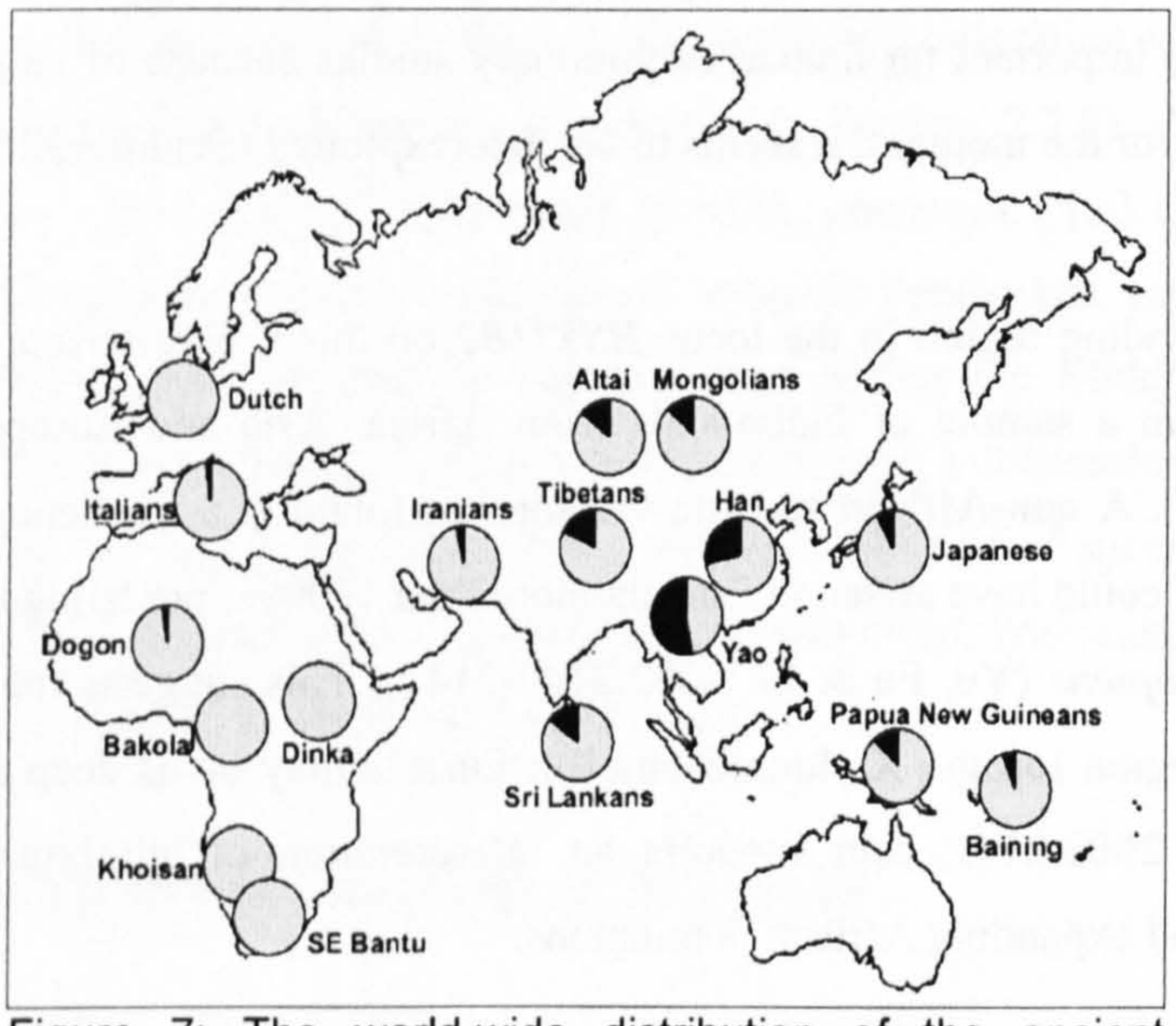

Figure 7: The world-wide distribution of the ancient *RRM2P4* lineage.

Reproduced from Garrigan et al. (2005b:191). The gradient is centered in East Asia.

Taken together, these studies support each other in suggesting that the X chromosome has a different evolutionary history from other genetic systems, especially mtDNA and the Y

chromosome. The most probable scenario, given that the loci reviewed are not linked, is that the X chromosome as an entity admixed from local archaic populations into the expanding modern *Homo sapiens*. The date of split of the independent admixing branches could go back to the initial spread of *Homo erectus* from Africa, ~2mya, or could be more recent, given recurrent gene flow between Africa and Eurasia. Another important implication is that it is also possible that coding, non-neutral genes also introgressed from the local archaic populations into the modern gene pool, providing a base for morphological and behavioral regional continuity. It seems that the X chromosome is particularly sensitive to introgression (Patterson *et al.*, 2006; Mallet, 2005; Payseur *et al.*, 2004; Payseur *et al.*, 2005), especially when differences in F_1 fertility between sexes appear, offering a plausible explanation for these findings.

A very recent paper (Evans *et al.*, 2006) presents a possible case of introgression involving a phenotypically important gene. More exactly, the gene *Microcephalin* (see Section 4.2 for details) is involved in brain growth and development (Evans *et al.*, 2006; Evans *et al.*, 2005; Gilbert *et al.*, 2005) and presents a haplogroup (named "derived", denoted "D") which is very recent, shows signs of natural selection and has marked geographic structure (Section 4.2; Evans *et al.*, 2005). It is argued that the MRCA of the D chromosomes dates to ~37kya, the MRCA of the non-D chromosomes dates to ~0.99mya and that "the D and non-D chromosomes belong to two distinct, deeply divided clades connected by a single branch around the root of the tree" (Evans *et al.*, 2006:2), coalescing at the much older age of ~1.7mya (Evans *et al.*, 2006:2; see Figure 8).

If reproductive isolation between these two branches is assumed, a separation time of ~1.1my results, but this could be much longer (but less than ~1.7my) if there was gene flow between these populations (Evans *et al.*, 2006:5). The most probable scenario (Evans *et al.*, 2006:5) seems to imply that the branch leading to modern humans fixed the non-D haplogroups while the other lineage fixed the D haplogroup and during an interbreeding event ~37kya the D haplogroup passed into the lineage leading to modern humans where, under strong natural selection, reached very quickly a global frequency of 70% (and probably still increasing). Given the geographic distribution of the D haplogroup, it is plausible that this introgression involved an Eurasian *Homo* lineage, the authors suggesting the Neanderthals as a candidate (Evans *et al.*, 2006:5), but one cannot rule out the Asian

Homo erectus.

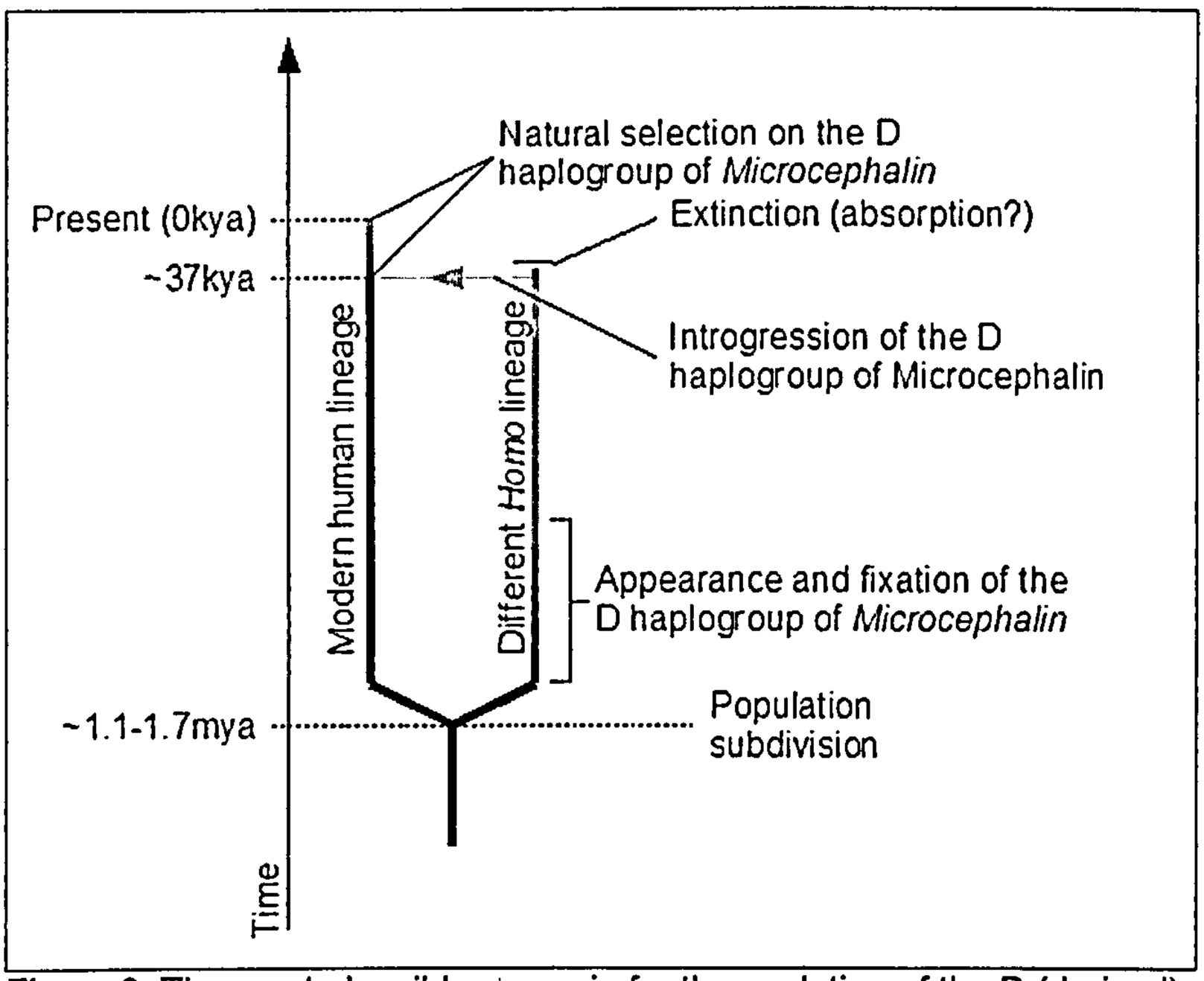

Figure 8: The most plausible scenario for the evolution of the D (derived) haplogroup of Microcephalin.

Redrawn after Evans et al. (2006). Shades of gray represent the frequency of the D haplogroup (black = 0%, very light gray = 100%).

But irrespective of the exact nature of the *Homo* variant from which the D haplogroup of *Microcephalin* introgressed into the lineage leading to modern humans, it very strongly rejects the separate species status for modern humans and suggests that some highly adaptive characters actually originated in other *Homo* lineages and accreted into the modern phenotype. More speculatively, given the role in brain growth and development played by *Microcephalin*, one can start wondering about the real cognitive capacities of "archaic" *Homo*. Also, it is possible that other selectively non-neutral genes have introgressed in the modern gene pool.

2.2.4. Primate models and the speciosity of *Homo*

Inferences about the number of fossil species are notoriously difficult (Relethford, 2003:46-50; Relethford, 2001:51-54; Tattersall & Mowbray, 2004:377) and almost always involve comparison with living models. In the case of fossil hominins, the predilect model was

represented by living primates, usually the extant humans and the great apes (*Pan troglodytes* and *P. paniscus*) (Jolly, 2001:177; Cameron, 2003:3; Villmoare, 2005:4; Tattersall & Mowbray, 2005:376-377). The conclusion from such comparisons seems to be that the morphological diversity of fossil *Homo* exceeds the intra-species diversity in living humans and great apes (Tattersall & Mowbray, 2004:377; Harvati, Frost & McNulty, 2004) and, thus, the speciose model for *Homo* is adequate (Cameron, 2003:26), but there are convincing arguments that the models used are not appropriate for the task (Jolly, 2001; Holliday, 2003; Hunt, 2003).

Papio (baboons) and *Theropithecus* (gelada) are usually classified as distinct genera (Holliday, 2003:656; Section 2.1.2.1), as seemingly justified by their important morphological differences (Jolly, 2001; Holliday, 2003:656), and they diverged ~5mya (Jolly, 2001:189). Given these, it is surprising that they frequently hybridize both in nature (Holliday, 2003:657; Jolly, 2001:189) and in captivity (Jolly, 2001:189), and the hybrids seem to be both viable and fertile (Holliday, 2003:657; Jolly, 2001:189-190, 197): "[...] there is no evidence for hybrid breakdown, behavioral incompatibility, or intrinsic sterility" (Jolly, 2001:196).

Prompted by this evidence, Clifford Jolly defines *allotaxa* as "phylogenetically close, but well-differentiated and diagnosable, geographically replacing forms whose ranges do not overlap, but are either disjunct, adjoining or separated by comparatively narrow zones in which characters are clinally distributed" (Jolly, 2001:193-194), and argues that *Papio* and *Theropithecus* are both allotaxa (Jolly, 2001:196). Trenton Holliday (2003:657) further analyzes this concept and concludes that allotaxa belong to the same biological species (BSC), but probably to different phylogenetic species (PSC). Botanists have long known that morphological differences and a long history of separate evolution do not automatically equate biological species status, and as early as the beginning of the 20th century the notion of *syngameon* was defined, as a set of closely related species of plants which commonly hybridize (Lotsy, 1925, cited in Holliday, 2003:656): "plant taxonomists frequently group species in larger units called *syngameons*, within which natural hybridization may take place[, y]et the species within a syngameon remain separate species" (Skelton, 1993:375). These two notions, allotaxa and syngameon, represent the same reality, whereby apparently distinct groups of organisms prove to belong to the same species. Another related concept is

represented by *polytypic species* (West-Eberhard, 2003:378), but seems to be more circumscribed.

If we use these living primates as models for extinct hominins, as convincingly argued by Jolly (2001), Holliday (2003) and Hunt (2003), then we are faced with the conclusion that, possibly, the genus *Homo* was *not* speciose at all, forming a single syngameon, with the hominins being fully interfertile allotaxa over their geographical and temporal range. This "[...] suggests that all human lineages stemming from the *H. ergaster* stock were probably as fully interfertile as are extant *Papio* populations. On these grounds, they could be regarded as members of a single, polytypic (BSC) species" (Jolly, 2001:196) and "[a] strict papionin analogy would therefore argue that all *Homo* (*sensu stricto*) were interfertile" (Holliday, 2003:659), and:

> [l]iving primates by no means suggest that the hominin clade should be speciose; rather, they suggest the opposite. Arguably, no two contemporaneous hominin species were separated by significantly more time, 1.6 Ma, than has separated the demonstrably single-species chimpanzee. Hominin "species" distinctness might have been maintained more by allopatry or centripetal niche separation than reproductive isolation (Hunt, 2003:499).

Of course, one might argue that having a single isolated case of primates forming a syngameon, doesn't necessarily entail that extinct *Homo* behaved in such a "deviant" way, but, as Jolly (2001), Holliday (2003) and, especially, Hunt (2003) argue, this is not a single, awkward case (*Pan troglodytes* and vervets are also considered by Hunt, 2003). Moreover, Hunt (2003) argues that the usual speciose primate models for extinct *Homo*, *Cercopithecus* and *Macaca* are unfit for the task:

> Cercopithecus is a very unusual genus. Its small body size, arboreality, relatively small brain, small home range, forest habitat and low sexual dimorphism argue against it as a good analog for hominin species richness. Largely for these reasons and because it is endemic to Asia, Macaca is an even poorer analog (Hunt, 2003:489)

while "[o]ur closest relatives and baboons exhibit typical primate speciosity. We should expect typical speciosity among hominins as well. We should expect two species per genus." (Hunt, 2003:489).

For a long time, hybridization and introgression were regarded as infrequent and somehow "deviant" by zoologists, as opposed to botanists, mainly due to Mayr's conception (BSC),

but, during the last decades, its frequency and importance started to be reconsidered (Mallet, 2005; Dowling & Secor, 1997; Bull, 2003; Seehausen, 2004; Arnold *et al.*, 2001). The basic assumptions of hybrid breakdown and infertility are questioned (Seehausen, 2004; Arnold *et al.*, 2001) and molecular techniques allow better assessment of its incidence in nature (Bull, 2003; Mallet, 2005). For example, Mallet (2005:231, Table 1) lists the percent of hybridizing species (e.g., 6.5-14.8% for Swallowtail butterflies, 26% for Passion flower butterflies, 9.3% for world birds and 6.0% for European mammals) and concludes that hybridization and introgression are much more frequent than generally assumed. Studies of particular cases (e.g., Bull, 2003, for *Heliconius*; Donnelly *et al.*, 2004, for *Anopheles gambiae*; Saltzburger, Baric & Sturmbauer, 2001, for *Neolamprologus marunguensis*; Ranganath & Aruna, 2003, for *Drosophila nasuta* and *D. albomicans*; Tranah, Campton & May, 2004, for *Scaphirhynchus albus* and *S. platorhynchus* and Young *et al.*, 2001, for *Oncorhynchus mykiss irideus* and *O. clarki clarki*) also suggest that their frequency and evolutionary importance were underestimated, leading to increase in biodiversity and inter-specific gene flow. Therefore, the case of *Papio* and *Theropithecus* is not rare nor deviant.

Support for this theory is also provided by a recent study (Patterson *et al.*, 2006) of the divergence times across the genomes of humans and chimpanzees. As opposed to the standard average divergence estimates, the genetic divergence of specific regions vary between 0.84 and 1.47 times this average, representing a range of ~4my (Patterson *et al.*, 2006:1106), with the X chromosome the last to diverge (Patterson *et al.*, 2006:1105). This suggests that "[...] the hominin and chimpanzee lineages initially separated but then exchanged genes before finally separating less than 6.3 Myr ago" (Patterson *et al.*, 2006:1106). Therefore, this speciation turns out to be far from an ideal, punctual event and more akin to a reticulate process, whereby genes continue to be exchanged between allotaxa for a long time (millions of years) after their "separation".

In conclusion, the arguments for an appropriate model of living primates seem to support a low species count view for *Homo*, where the different geographical and temporal forms (allotaxa) belong to the same biological species. This does not assume, of course, a panmictic, homogeneous population of *Homo* throughout the Plio-Pleistocene Old World and does not imply the non-existence of regional characteristics and continuity. "[...] [T]he assumption of universal interfertility within the genus *Homo* (strictu sensu) [does not]

conflict with evidence pointing to long-term, consistently diagnosable human lineages [...]" (Jolly, 2001:196).

2.2.5. Regional morphological continuity

Regional continuity can be defined as the persistence for long periods of evolutionary time of specific (combinations of) *features*[79] in a given geographical region (Wolpoff & Caspari, 1997:293; Relethford, 2001:57; Lewin, 1998:393). In the context of human evolution, regional continuity *across* assumed species boundaries is very important, as it could, in principle, decide between replacement versus admixture models. More specifically, if such features (or combinations of features) could be identified, crossing the putative boundary of modern *Homo sapiens* expansion out of Africa and replacement of the archaic populations, then a good argument against replacement without admixture can be constructed. The main issue in interpreting such a case of regional continuity concerns the (combinations of) features' relevance for the demographic history of the region as opposed to its evolutionary history (Wolpoff & Caspari, 1997:293-299; Relethford, 2001:57-59).

Natural selection on a given trait can determine parallel or convergent evolution[80], thus mimicking regional continuity. For example, consider skin color in human population: it tends to be darker the closer to the equator the population is located, an effect which seems to be due to natural selection in relation to UV radiation (Jobling, Hurles & Tyler-Smith, 2004:413-416). Thus, for any given geographical region, there is an optimal range of skin colors towards which the inhabiting populations tend to converge, this range representing a case of regional continuity non-informative to the demographic history of the region. Suppose a new population arrives and either replaces or admixes with the local population: after a short evolutionary time (Relethford 2003), the outcomes will be indistinguishable. A similar example is represented by body structure as an adaptation to climate (Relethford, 2003; Jobling, Hurles & Tyler-Smith, 2004:403).

79 As opposed to the persistence of "races", with which is usually confused. This last view forms the basis of polygenic models of human evolution (Wolpoff & Caspari, 1997).

80 *Convergent* evolution is defined as the evolution of the considered lineages towards greater similarity (the distance between them decreases) (Skelton, 1993:751), while parallel evolution involves some closely related lineages which evolve in a similar fashion (as opposed to diverging or converging), the distance between them remaining constant (Skelton, 1993:753).

The most informative traits, in this respect, must be selectively neutral. A classic example is offered by the *mandibular foramen* (Wolpoff & Caspari, 1997:296-297; Relethford, 2003:99): this is an opening on the internal face of the mandibular ramus through which blood vessels and a nerve branch pass (Seeley, Stephens & Tate, 2005:217). It is polymorphic in humans and has two main shapes: *horizontal-oval* and *normal* – the most frequent form in living populations (Wolpoff & Caspari, 1997:297). The two forms are presumably selectively neutral (Wolpoff & Caspari, 1997:297; Relethford, 2003:99; Relethford, 2001:204) and concerning their distribution across space and time:

> [t]he horizontal-oval mandibular foramen is virtually unique to European fossils. It is found in almost no other remains [...] [,b]ut the horizontal-oval foramen has a significant frequency in the subsequent post-Neandertal populations of Europe and only decreases to rarity in recent Europeans (Wolpoff & Caspari, 1997:297).

The following table summarizes the temporal pattern of distribution of the mandibular foramen forms in European populations:

Population	***Horizontal-oval Freq (%)***	***Normal Freq (%)***
Neandertal	53	47
Early Upper Paleolithic	18	82
Late Upper Paleolithic	7	93
Mesolithic	2	98
Medieval	1	99

Table 1: The distribution of the mandibular foramen polymorphisms across time in European population .

Adapted from Wolpoff & Caspari, 1997: 297.

A trend towards fixation of the *normal* variant seems to be present in the European population through evolutionary time, reaching almost complete fixation in the modern sample. But the most important conclusion to be drawn from this example is that this putatively selectively neutral feature has crossed the Neanderthal-modern *Homo sapiens* boundary, pointing towards admixture between the incoming moderns and preexisting archaics. This trend towards fixation can be explained through genetic drift and admixture between two unequal populations, the modern *Homo sapiens* being more numerous (Relethford, 2003:93-99; Relethford, 2001:202-205).

The general problems faced by studies addressing regional continuity across human

evolutionary periods, especially the issue of the emergence of the modern *Homo sapiens*, are represented by the nature of the available material: fossils. This restricts the range of usable features or combinations of features to morphological characteristics and also vastly reduces the sample sizes. The features or combinations of features under study must be present in all the analyzed fossils, but given the characteristics of the taphonomic processes (Skelton, 1993:564-576) and the vagaries of fossil discovery and availability, the samples are very small. Another potential problem is represented by the possible non-independence of the considered traits, which, compounded with the small sample sizes, could alter the statistical significance of the results. But there are a number of studies (besides the mandibular foramen case mentioned above) which address these problems and report cases of regional morphological continuity.

Wolpoff, Hawks, Frayer & Hunley (2001) analyzed transitional forms (crania) from two peripheral regions (Australia: Willandra Lakes Hominid 50 and Czech Republic: Mladeč 5 and 6) and, after a pairwise comparison, concluded that they have a dual ancestry, invading modern and local archaic:

> [w]e do not doubt that many prehistoric groups were replaced by other, but we conclude that the hypothesis that all living humans descended from a single geographically isolated group during the Late Pleistocene is false, and that the replacement explanation for the origin of these early modern Australians and Europeans can be ruled out (Wolpoff, Hawks, Frayer & Hunley, 2001:296).

Demeter, Manni & Coppens (2003) used a morphometric analysis of 45 fossil crania from the Far East (Demeter, Manni & Coppens, 2003:627). They support regional continuity for this area and conclude that

> [t]he 2 major morphologies [...] described in this work illustrate the coexistence of at least 2 well characterized types of the first modern human groups that colonized the Far East during the Late Upper Pleistocene, validating the multi-regional evolution hypothesis theory (Demeter, Manni & Coppens, 2003:637).

The origin of moderns in China is hotly debated and Wu (2003) shows that culturally and skeletally, there was evolutionary continuity between *Homo sapiens sapiens* and earlier *Homo sapiens erectus*[81]. He proposes a theory of continuity with hybridization (Wu, 2003:134), which is essentially a variant of multiregionalism. The oldest European modern human (to date) was discovered in Romania, Peştera cu Oase (Trinkaus *et al.*, 2003), and dated to 34-36kya: this mandible presents a "mosaic of archaic, early modern human and

81 Even the varieties names suggest the author's conception that the two belong to the same species.

possibly Neandertal morphological features" (Trinkaus et al., 2003), suggesting, at least, that modern humans continued to evolve after leaving Africa, and, possibly, that they did interbred with local archaics and that regional features persisted into the modern populations.

Wolpoff *et al.* (2004) argue that

> [t]he supposedly unique Neandertal [morphological] features, such as the retromolar space [...], posterior placement of the mandibular mental foramen [...], taurodontism [...], the lateral (in contrast to the superior) frontal sinus conformation [...], mastoid tubercle [...], suprainiac fossa [...], lambdoidal flattening [...], H-O mandibular foramen [...], dorsal axillary border configuration of the scapula [...], all show considerable variation within the Neandertals and a continuous distribution from Mousterian to early Upper Paleolithic populations (Wolpoff *et al.*, 2004:531)

and plausibly address the critics (Wolpoff *et al.*, 2004:531-533), concluding:

> [w]here are these Neandertal features today? The answer is that some have disappeared while others remain in Europe, and some of these are commonly used in forensic applications for determining ancestral affinities. [...] Included in these are [...]: (1) the high nasal angle involving the slope of the lofty nasal bridge, as it rises up between the orbits, incorporating the frontal processes of the maxillae as well as the nasal bones themselves; (2) the course of the zygomaxillary suture (turning inward at its inferior aspect); (3) the maxillary expansion at the lateral nasal borders; and (4) the lateral zygomatic orientation [...]. These features, and others like them, are not present in Neandertal contemporaries, such as those from Qafzeh in Western Asia or the Herto Ethiopian (Wolpoff *et al.*, 2004:533 and Plate 1:530).

These cases of morphological continuity between European Neanderthals and early modern humans, if confirmed, would forcefully argue for admixture in the origin of modern Europeans.

The issue of morphological regional continuity has a very long history (Weidenreich, 1947a, 1947b; Wolpoff & Caspari, 1997). It must be highlighted that this applies to specific cases and does not claim to be valid everywhere, as sometimes wrongly presented (e.g. Lewin, 1998:389). For example, even if the case of regional continuity from Neanderthals into modern Europeans would prove false, it would not automatically invalidate claims of regional continuity in East Asia. Given the sample of works cited above, it seems plausible, for the moment, to accept suggestions of morphological regional continuity. Moreover, given the possibility that the members of *Homo* were not different species and that they admixed, it is plausible that regional continuity might manifest in features unable to fossilize (soft-tissue, behavior) and thus, impossible to ascertain by studying the fossil record

2.2.5.1. The *Abrigo do Lagar Velho* child

Probably the best-known case of regional continuity is represented by the Abrigo do Lagar Velho child (Duarte *et al.*, 1999). It represents a largely complete skeleton of a ~4 years old child, discovered in the Abrigo do Lagar Velho, Lapedo Valley, central Portugal and dated at ~24ky. Probably the best reference for this important find is *Portrait of the Artist As a Child: The Gravettian Human Skeleton From the Abrigo Do Lagar Velho and its Archaeological Context*, Oxbow Books Ltd., 2003, edited by J. Zilhão and E. Trinkaus.

Morphologically, the child seems to be a hybrid between modern *Homo sapiens* and Neanderthals (Duarte *et al.*, 1999; Trinkaus & Zilhão, 2003), despite criticisms claiming it to be just a more robust modern (Tattersall & Schwartz, 1999). The mosaic is re-analysed by Erik Trinkaus and João Zilhão (Trinkaus & Zilhão, 2003:507-512) and they conclude (in concordance with the original analysis) that "[...] the nature of the mosaic for several complexes suggests an unusual combination of its ancestry" (p. 512) and that the mosaic is real and not just an illusion: "[...] it is apparent that the mosaic is real [...] [and] the mosaic is sufficiently documented not to be wished away" (p. 512), in response to the acid earlier critics of Ian Tattersall and Jeffrey Schwartz (1999), which conclude that

> [...] the analysis of Duarte *et al.* of the Lagar Velho child's skeleton is a *brave and imaginative interpretation* [...] the specimen itself lacks not only derived Neanderthal characters but *any suggestion of Neanderthal morphology*. The probability must thus remain that this is *simply a chunky Gravettian child* [...] (Tattersall & Schwartz, 1999:7119, *italics* mine).

It seems that the hybrid is real and, concerning its ancestry, Trinkaus & Zilhão conclude that "Lagar Velho I is therefore extremely unlikely to be an individual randomly sampled from a representative European Gravettian early modern human population [...] also extremely unlikely that this individual represents a normal Neandertal [...] [t]he admixture hypothesis therefore stands" (Trinkaus & Zilhão, 2003a:513-514).

This Neanderthal-modern human hybrid is very important for the issue of regional continuity and the status of archaic *Homo*, because it originates from a geographical region (Portugal) and period (~25kya) when the last Neanderthals and modern humans coexisted in Europe for a prolonged interval, thus maximizing both the probability of admixture and the probability of fossilization of such hybrids (Trinkaus & Zilhão, 2003). It must be highlighted that the window of opportunity for such recognizable hybrids to fossilize is very small, because of

the inequality of the two admixing gene pools and the relatively swift swamping of Neanderthal genes by incoming, modern human African genes (Relethford, 2001; Trinkaus & Zilhão, 2003). The specific issue of the fertility of this type of hybrids cannot be directly addressed for the moment, because of lack of fossils showing different degrees of admixture, and, from the nature of the Lagar Velho hybrid it cannot be asserted if it does represent an F_1 individual or not (Trinkaus & Zilhão, 2003:517).

Another critical point concerning this hybrid is represented by its social acceptance. Its burial context, as analysed in the larger context of Middle and Early Upper Palaeolithic burials (Zilhão & Trinkaus, 2003a), suggests that this individual was recognized as a full member of the community and not perceived as some kind of freak, resulting from the unnatural mating between man and beast. This, in turn, supports the view that such admixture was considered at least tolerable, and it was probably frequent enough to gain social acceptance. Moreover, it also supports the hypothesis that modern humans and Neanderthals regarded each other as humans, contra pervading academic (e.g., Stringer & McKie, 1996) and literary (e.g., Baxter, 2003) claims to the contrary.

"The broader implication of Lagar Velho I is a final rejection of the Late Pleistocene Out-of-Africa with complete replacement scenario for modern human emergence" (Trinkaus & Zilhão, 2003a:516) and forcefully suggests that introgression from local archaic human populations into the modern humans is a real possibility.

2.2.6. Global trends

The fossil record seems to show some common evolutionary trends over the entire geographic range of *Homo* (Relethford, 2001:58), which, if confirmed, would be hard to explain by a model assuming a number of different biological species, where the only valid evolutionary explanation would be parallel evolution. But if *Homo* is composed of many allotaxa connected by gene flow, the alternative explanation of the spread of characters is available (Relethford, 2001:58; Wolpoff & Caspari, 1997:270-313).

Probably the best known such trend is represented by the increase in brain size (Lee & Wolpoff, 2003): "[b]rain size increase is unarguably one of the most distinct and significant

evolutionary trends in Pleistocene human evolution" (Lee & Wolpoff, 2003:186). They try to discriminate between competing models: gradualism and continuity versus stasis in some human lineages versus different rates in different regions (Lee & Wolpoff, 2003:186). 94 cranial capacities of fossils living between 50kya and 1.8mya (Lee & Wolpoff, 2003:189) were used, analyzing trends in the log-log transformation of cranial capacity versus time (Lee & Wolpoff, 2003:189-191). The results support a single evolutionary process of increasing brain size, incompatible with a punctuated pattern (Lee & Wolpoff, 2003:191) and also suggest that the same process accounts for earlier and later data (Lee & Wolpoff, 2003:193):

> [g]radual change in cranial capacity, in the sense of temporal variation responding to a single underlying process, is compatible with the single lineage interpretation of Pleistocene *Homo* and difficult to reconcile with current speciose interpretations of Pleistocene human evolution (Lee & Wolpoff, 2003:194).

Another, more elusive trait, showing global trends is gracilization[82] (Wolpoff & Caspari, 1997:26).

The parallel evolutionary explanation of these global trends is certainly plausible, but less parsimonious than the alternative one, involving gene flow. For parallel evolution to work, common selective pressures must be identified, and, in the case of increase in brain size, a sort of positive feedback between culture and cognitive capacity could be invoked, proposing that once culture entered the scene, there is no turning back but a steady pressure for increased cognitive capacity capable of dealing better with culture, generating, thus, more complex cultures in turn. A similar explanation could be advanced for gracilization, seen as a side-effect of reliance on culture and relaxation of the selective pressures on brute force. The alternative explanations, involving gene flow, also appeals in some cases to common selective pressures favoring the spread of certain alleles, conferring global selective advantages, but deals better with the temporal synchronization of the trends across vast geographical spaces and putative species boundaries. Without this synchronizing network of sharing advantageous alleles, one has to postulate intrinsic factors explaining the apparent sameness of rate across species, but such a convincing mechanism does not seem to have been proposed. The overall conclusion is that if these synchronous, global trends are real, then a view of *Homo* as composed of allotaxa connected through constant gene flow becomes more probable than the alternative, multiple species, view.

82 Usually confused with "modernity" (Wolpoff & Caspari, 1997).

2.2.7. Ancient DNA

The literature dealing with the Neanderthals (Section 2.1.2.3), both academic, popularization and fiction, is so vast that an exhaustive review is utterly impossible. The claims range from New Ageist absurdity (e.g. Darnton, 1996), through utter primitivism (e.g., Stringer & McKie, 1996), to full humanity (e.g., Trinkaus & Zilhão, 2003). My own point of view is that the similitudes with modern humans, both behavioral and anatomical, are overwhelmingly more important than the differences. I think that if the fantastic dream of having a living Neanderthal (or, better, a natural community) transported into the present (Asimov & Silverberg, 1993), they would probably integrate into the western culture as well (or as badly) as any traditional modern human culture (e.g., New Guinea or the Amazonian basin), and would probably provide not much material for the study of any "alternative ways of thinking".

In 1997, a new chapter of the controversy concerning the modern human origins was opened by a seminal paper (Krings *et al.*, 1997), reporting the first successful extraction of ancient DNA from the Neanderthal-type specimen. The difficulties facing this type of studies cannot be overstated and include contamination with modern DNA and decay through time (Relethford, 2003:80-84; Jobling, Hurles & Tyler-Smith, 2004:110-113, Box 4.5:115-116, Box 4.7:117-118). Therefore, analyses of ancient DNA must obey strict protocols and meet high standards of quality in order to be accepted as valid. Moreover, only mtDNA can be successfully extracted and analyzed with present technology, given that it is much more abundant than nuclear DNA. Since this first paper, many other successful extractions of DNA from Neanderthal and early modern human specimens were performed.

Serre *et al.* (2004) compared ancient mtDNA extracted from 4 Neanderthals (Vindija 77 – Croatia, Vindija 80 – Croatia, Engis 2 – Belgium and La Chapelle-aux-Saints – France) with 5 early modern humans (Mladeč 25c – Czech Republic, Mladeč 2 – Czech Republic, Cro-Magnon – France, Abri Pataud – France and La Madeleine – France); Caramelli *et al.* (2003) extracted mtDNA from two early modern humans from the Paglicci cave in southern Italy and compared them with already extracted Neanderthal sequences; Lalueza-Fox *et al.* (2005) extracted mtDNA from a Neanderthal specimen from El Sidrón Cave, Asturia, North Spain (El Sidrón 441 tooth); Ovchinnikov *et al.* (2000) extracted mtDNA from a specimen from

the Mezmaiskaya cave in the northern Caucasus, one of the easternmost Neanderthal populations, Krings *et al.* (1999) extracted from the type specimen another sequence (HVRII) and Krings *et al.* (2000) extracted mtDNA from the Vindija 75 Neanderthal specimen in Croatia. In total, there are to date 9 ancient mtDNA sequences extracted from Neanderthal remains (Lalueza-Fox et al., 2005:1077). The overall pattern seems to be that Neanderthal mtDNA is different both from living modern and contemporary early modern humans (Jobling, Hurles & Tyler-Smith, 2004:260-262; Relethford, 2003:80-84; Krings *et al.*, 1997:25-26; Lalueza-Fox *et al.*, 2005:1079-1080; Caramelli *et al.*, 2003:6595; Serre *et al.*, 2004:0315; Ovchinnikov *et al.*, 2000:491-492; Krings *et al.*, 2000:145; Weaver & Roseman, 2005:680), but see Gutiérrez *et al.* (2002) for a critique; they argue that "the phylogenetic position of the ancient DNA sequences recovered from Neanderthal bones is sensitive to the phylogenetic methods employed [and] it depends on the model of nucleotide substitution, the branch support method and the set of data used" (Gutiérrez *et al.*, 2002:1363) and they even obtain a tree including Neanderthals inside the modern human clade (Gutiérrez *et al.*, 2002:1362, Figure 2B): "we believe that the likelihood mapping values supporting Neandertals as a different species might be artificially increased" (Gutiérrez *et al.*, 2002:1363). From a population-internal point of view, it seems the genetic diversity of the Neanderthals was comparable to that of modern humans (Jobling, Hurles & Tyler-Smith, 2004:260-261; Krings *et al.*, 2000:144; Ovchinnikov *et al.*, 2000:492; Lalueza-Fox *et al.*, 2005:1079), prompting Lalueza-Fox *et al.* (2005) to conclude that "[this] could suggest that the evolutionary history of Neandertals and modern humans were characterized by similar demographic parameters" (Lalueza-Fox *et al.*, 2005:1079).

The most important problem now is to establish *how* different the Neanderthal and modern human (living populations, but more relevantly, early modern human fossils) mtDNAs are (Jobling, Hurles & Tyler-Smith, 2004:261-262; Relethford, 2003:84-87): do the differences allow one to draw conclusions concerning the species status of the Neanderthals or their genetic contribution to living human populations?

The original report (Kings *et al.*, 1997) computes the pairwise differences between living humans, Neanderthal and chimpanzee mtDNA sequences and displays them (Kings *et al.*, 1997:25, Figure 6) in a depiction which probably became emblematic for the popular perception of ancient Neanderthal DNA and its significance:

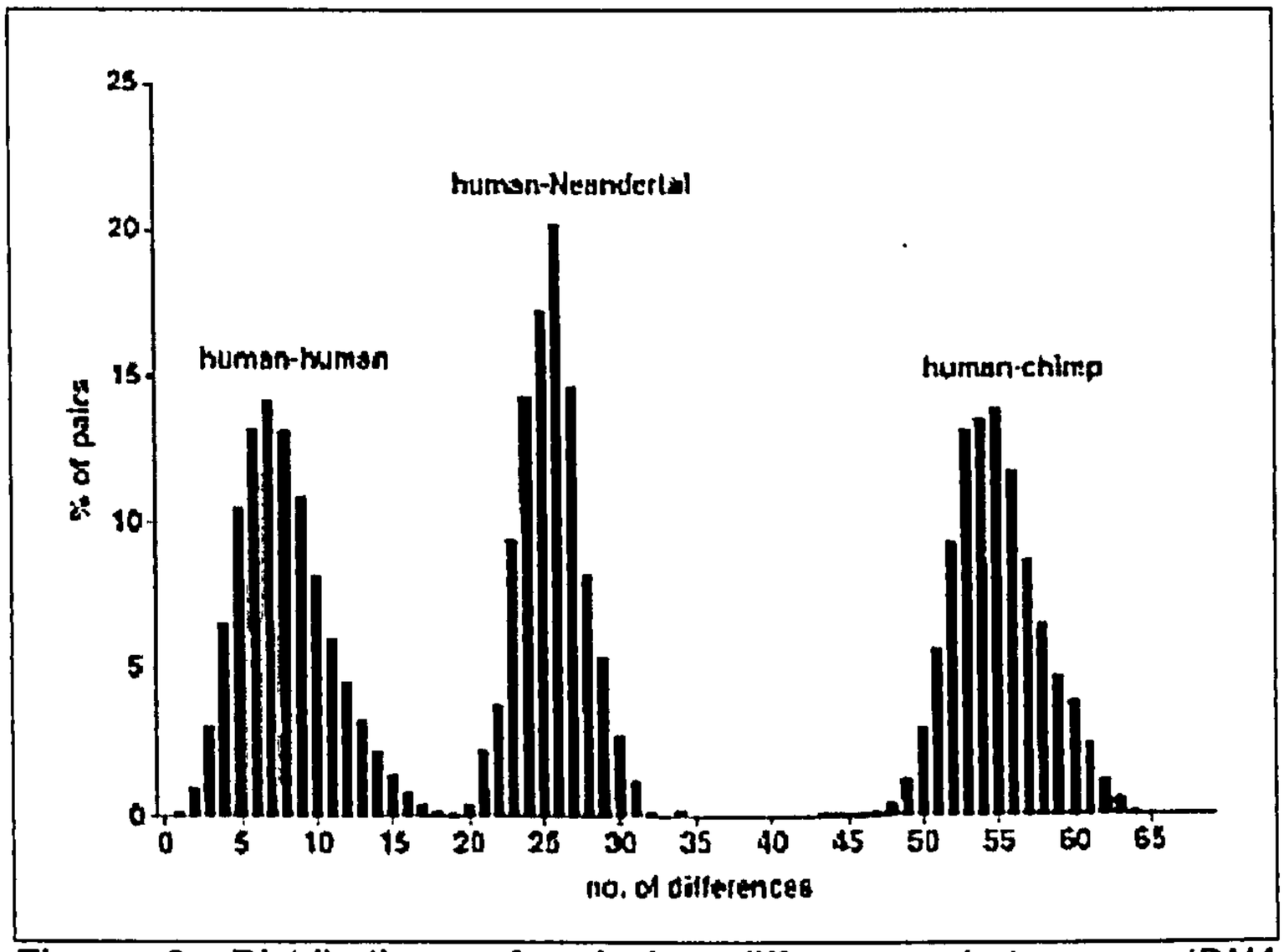

Figure 9: Distributions of pairwise differences between *mtDNA* sequences of living humans, living chimps and the original ancient Neanderthal *mtDNA* extraction.

Reproduced from Krings et al., 1997:25. Horizontal axis: the number of differences between sequences; the vertical axis: the percent of pairs showing that number of differences.

It can be seen that the human-Neanderthal comparison is outside the range of human-human distribution, suggesting that Neanderthal mtDNA (at least, the specific sequence analyzed) is outside the mtDNA pool of living humans (Krings *et al.*, 1997:24-25). Relethford (2001:190; 2003:88), reports a comparison between the original mtDNA Neanderthal sequence (Krings *et al.*, 1999) versus living humans and three chimpanzee subspecies (Western, Central and Eastern):

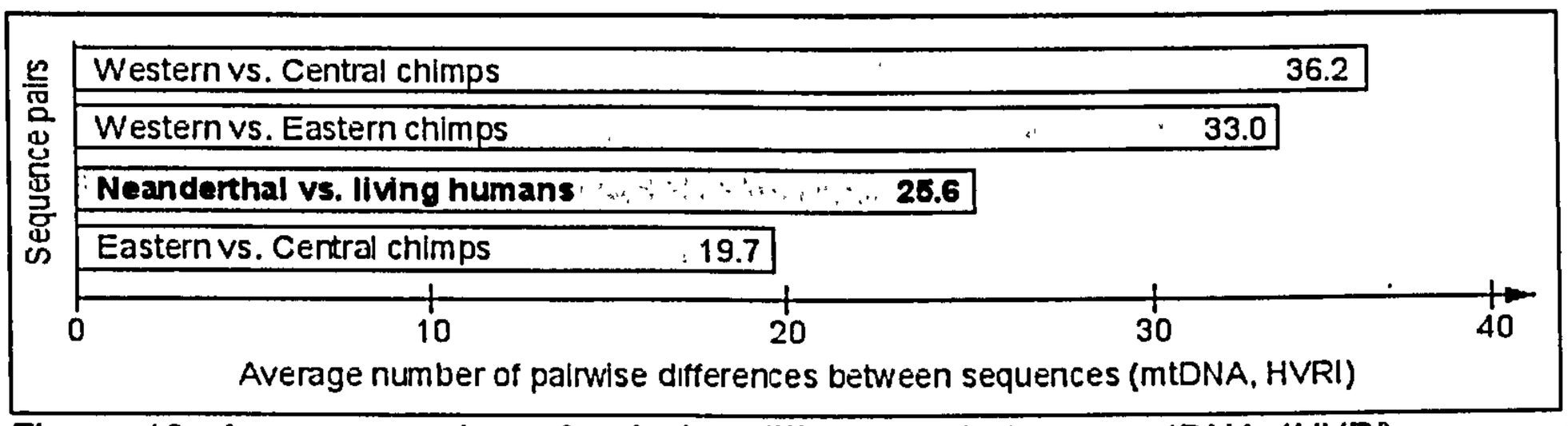

Figure 10: Average number of pairwise differences between *mtDNA* (*HVRI*) sequences compared between pairs of populations.

Considered populations: the Neanderthal type specimen, living humans and three living subspecies of chimps (western, central and eastern). Adapted from Relethford (2001:191; 2003:88).

and shows that the Neanderthal-living humans distance falls well within the range of intraspecies distances for living chimpanzees. Thus,

> [t]hese findings suggest that Neandertals and living humans *could* have belong to different subspecies within the same species, especially if we consider that the chimpanzee comparisons are all made at one point in evolutionary time (the present), whereas the Neandertal-living human comparison encompasses between 35,000 and 70,000 years, depending on the exact date of the Feldhofer specimen (Relethford, 2001:190, *italics* in original).

Another claim against admixture between Neanderthals and modern humans is sometimes adduced: if there has been indeed such admixture, then the mtDNA of Neanderthals should be closer to the modern (living and fossil) Europeans than to other populations, but the actual data show no such pattern (Serre *et al.*, 2004; Caramelli *et al.*, 2003; Relethford, 2001:186-187). The lack of such closeness between Neanderthals and modern Europeans as compared to other populations can be explained by gene flow between two unequal populations (Relethford, 2001:187-190; Relethford, 2003:91-95).

Nevertheless, the actual Neanderthal sequences extracted from all the specimens so far, are *missing* from living human populations, suggesting to many that they were a separate species, which did not contribute genetically to the current gene pool, being totally replaced by the incoming African moderns (Relethford, 2001:181-182; Relethford, 2003:83-84; Jobling, Hurles & Tyler-Smith, 2004:261). The estimated TMRCA for the Neanderthal and modern human mtDNA sequences is estimated at 365-853kya (Ovchinnikov *et al.*, 2000:492), predating the emergence of archaic modern *Homo sapiens* in Africa, suggesting, again, that the Neanderthals were a separate evolutionary branch of *Homo*. But this conclusion proves to be unwarranted by the present data.

Mathematical biologist Magnus Nordborg (Nordborg, 1998; 2004) has argued that mtDNA alone (nor, for that matter, any other single genetic locus) cannot differentiate between a Neanderthal genetic contribution to modern populations and none at all. More exactly, a single genetic locus can reject only a perfectly panmictic population (Nordborg, 1998:1239) or the hypothesis of no interbreeding at all (if the sequences differ too little, Nordborg, 1998:1239), but it cannot support, nor reject, any more complex model of admixture (Nordborg, 1998; 2004). Thus, "[t]he fact remains that an inference about population properties that is based on a single locus (or a nonrecombining genome) is an inference from

a single data point" (Nordborg, 1998:1239), and, moreover,

> [...] if there has been a recent selective sweep in human mtDNA, even random mating cannot be rejected (Nordborg, 1998:1239).
>
> [...] [I]t is highly likely that even if Neanderthal mtDNAs existed among anatomically modern humans 50 KYA (say), they would all have been lost by now. [...] Further studies of mtDNA will tell us nothing: it is necessary to take a genomic approach. [...] [M]odern DNA contains very limited information about what happened in the past (Nordorg, 2004).

As is now largely recognized, the history of a locus is not necessarily the history of a population, and many independent loci are required for a convincing reconstruction of past demographic events (Relethford, 2001; 2003; Jobling, Hurles & Tyler-Smith, 2004). For example, Wall (2000) evaluates the number of necessary such loci at ~50-100 and concludes that, currently, there is not enough data to support either total replacement or admixture (Wall, 2000:1276-1278).

Further confirmation that the evolutionary history of different loci does not necessarily have to agree and illuminate the demographic history, came in 2001, with the publication of the analysis of mtDNA extracted from modern human fossils in Australia (Adcock *et al.*, 2001a). Ancient mtDNA was successfully extracted from 4 morphologically gracile individuals (Lake Mungo, southeastern Australia) and 6 robust individuals (Kow Swamp, northern Australia). LM3 is of Pleistocene age, dated at the time of the publication at 62±6kya, while the other three graciles, LM4, LM15 and LM55 are Holocene. The robusts were dated to the end of Pleistocene – the beginning of Holocene (Adcock *et al.*, 2001a:538). It must be noted that the robust morphologies are outside the range of living indigenous Australians, but it is generally agreed that they contributed to the modern populations (Adcock *et al.*, 2001a:538). The results suggest that the sequence of LM3 is outside the living human gene pool and partially survives only as a nuclear insert (Adcock *et al.*, 2001a:540-541), while the robust mtDNA is well within the current range of variation.

> Sequences from the lineage that includes LM3's mtDNA no longer occur in human populations, except as the nuclear Insert on chromosome 11. The fact that LM3's morphology is within the range of living indigenous Australians indicates that the lineages of the alleles contributing to this *gracile* phenotype have survived. In contrast, the mtDNAs of the *robust* KS individuals belong to the contemporary human lineage. Their distinct *robust* morphology has not survived intact, implying that the allelic lineages of many of the genes that contribute to this phenotype have been lost (Adcock *et al.*, 2001a:541, *italics* in original).

The suggested explanation is that a later selective sweep of the current mtDNA lineage replaced the mtDNA lineage of LM3 in Australia (Adcock *et al.*, 2001a:541). This finding clearly shows that simplistic interpretations of the ancient genetic data must be avoided. The methodology and results of this paper were contested by Cooper *et al.* (2001), but see the authors' response to these criticisms (Adcock *et al.*, 2001b). Later, the fossils (LM3) were redated to 40±2kya by Bowler *et al.* (2003), but this does not alter the main argument that the history of mtDNA can be decoupled from that of other nuclear loci and of the demographic history.

Very recently, preliminary results of an extremely ambitious project to sequence the entire Neanderthal genome have been published (see, for example, Pennisi, 2006), with one team unable to find any evidence of admixture, but the other team, lead by S. Pääbo (based in the Max Plank Institute for Evolutionary Anthropology in Leipzig), did find such evidence for a directional gene flow: "Taken at face value, our data can be explained by gene flow from modern humans into the Neandertals" (cited in Pennisi, 2006:1070). However, given the difficulties involved, it will be necessary to wait for a close scrutiny of the data and techniques used before drawing any firms conclusions. Moreover, it is possible that further results will invalidate these preliminary claims, again suggesting skepticism.

The overall conclusion from ancient DNA studies, so far, seems to be that mtDNA alone cannot discriminate between the competing models for the evolution of modern humans and that much more independent loci are needed. Given the technical difficulties involved, it is to be expected that a definitive answer based on this type of data will not be available in the near future.

2.2.8. The genetic structure of living populations

It is generally agreed that living humans are genetically very uniform (Jobling, Hurles & Tyler-Smith, 2004:250-252, 277-280; Relethford, 2001:101-104). For example, Alan Templeton (1998) compared F_{ST} for living humans and various other species of large-bodied mammals with excellent dispersal abilities (Figure 11). The genetic diversity of living humans, despite their worldwide distribution, is rather low, comparable to that of more circumscribed species (waterbuck, impala, wildebeest) and much lower than that of species

with comparable ranges (wolf) (Templeton, 1998:633; Jobling, Hurles & Tyler-Smith, 2004:278-279).

The same picture of high genetic uniformity seemed to also hold true when living humans were compared to their closest relatives, the chimpanzees: three times for the X chromosome and mtDNA and up to seven times for the Y chromosome (Yu *et al.*, 2003:1511; Harding & McVean, 2004:670), but turned up to be only approximately 1.5-2.0 times when autosomes where also considered (Harding & McVean, 2004:671; Yu *et al.*, 2003:1516-1517). Nevertheless, it seems safe to conclude that humans are a relatively genetically uniform species.

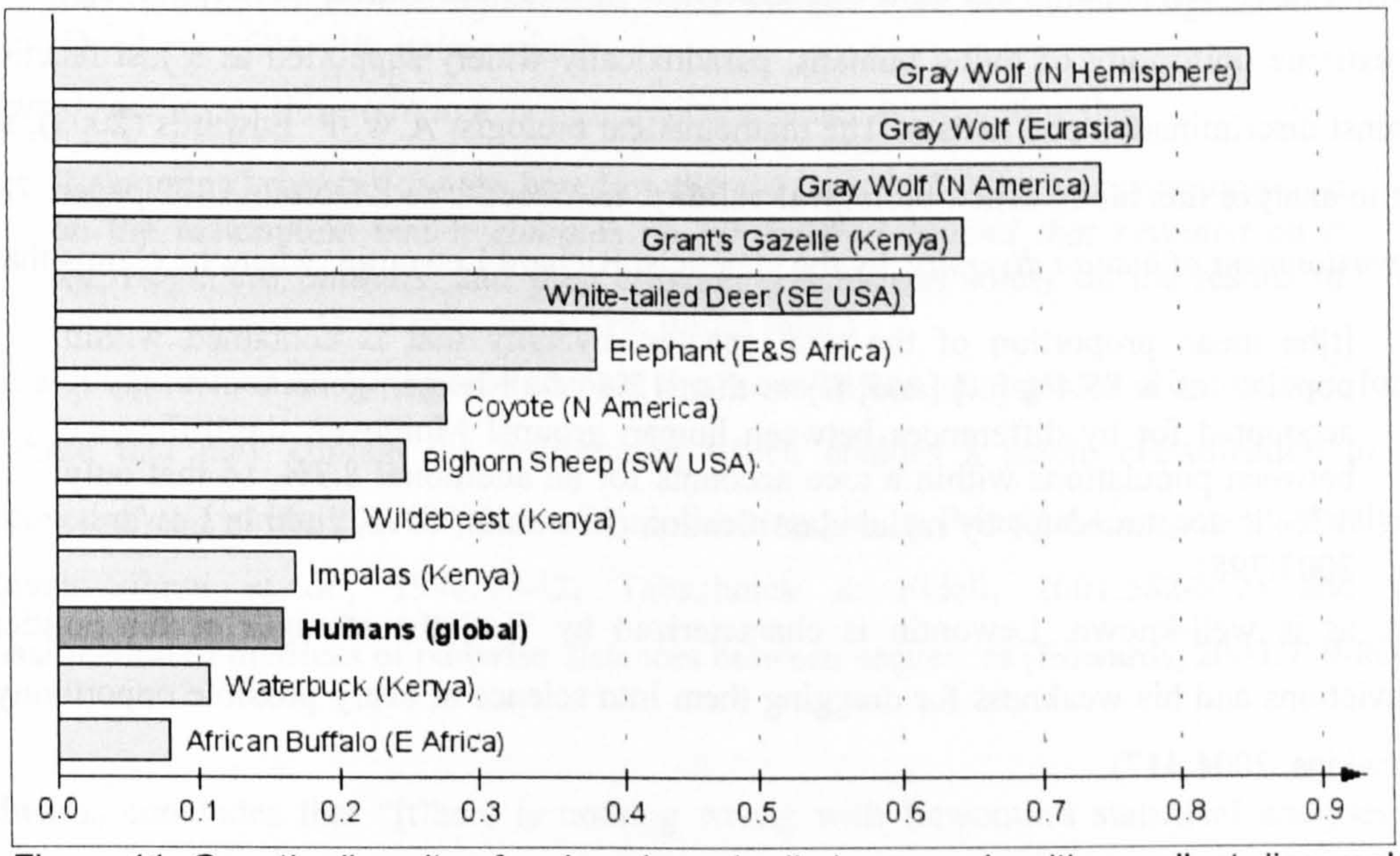

Figure 11: Genetic diversity of various large-bodied mammals with excellent dispersal abilities.

Adapted from Templeton, 1998:634. The horizontal axis represents F_{ST}.

The next interesting question concerns the apportionment of this genetic diversity across living humans (Jobling, Hurles & Tyler-Smith, 2004:277; Relethford, 2001:94).

2.2.8.1. The apportionment of genetic diversity in living humans and its interpretations

The current orthodoxy, popularized over and over again, states "usually without any reference that about 85% of the total genetic variation is due to individual differences within populations and only 15% to differences between populations or ethnic groups" (Edwards, 2003:798), suggesting that "the division of *Homo sapiens* into these groups is not justified by the genetic data [and that] [p]eople all over the world are much more similar genetically than appearances might suggest" (Edwards, 2003:798).

This seems to be one of those ideas which gain popularity simply by being repeated, as they match *l'esprit du temps*. In our case, this appears to be a hard scientific fact supporting views of extreme uniformity of living humans, paradoxically widely supported as a just reaction against discrimination and racism. The mathematical biologist A.W. F. Edwards (2003), set out to analyze this idea's actual biological validity: he traced it to a famous 1972 paper, *The apportionment of human diversity*, by the geneticist Richard Lewontin, where he claims that:

> [t]he mean proportion of the total species diversity that is contained within populations is 85.4% [...] [and] [l]ess than 15% of all human genetic diversity is accounted for by differences between human groups! Moreover, the difference between populations within a race accounts for an additional 8.3%, so that only 6.3% is accounted for by racial classification (Lewontin, 1972, cited in Edwards, 2003:798).

But, as is well-known, Lewontin is characterized by "[...] the strength of his political convictions and his weakness for dragging them into science at every possible opportunity" (Dawkins, 2004:417).

The study used the allele frequencies of 17 classical markers in 7 populations ("races") called in the paper: Caucasians, Black Africans, Mongoloids, South Asian Aborigines, Amerinds, Oceanians and Australian Aborigines and genetic diversity was measured by the Shannon information measure (somewhat similar to Nei's gene diversity) (Jobling, Hurles & Tyler-Smith, 2004:277). Later work largely confirmed these results (summarized in Jobling, Hurles & Tyler-Smith, 2004:278, Table 9.1): on average, autosomal variation is apportioned ~83-88% within populations and ~9-13% between (continental) populations, the notable exceptions being mtDNA and the Y chromosome (Jobling, Hurles & Tyler-Smith, 2004:278).

But what *is* arguable is the conclusion he draws from these data, conclusion accepted almost without comment by later work (see, for example, Jobling, Hurles & Tyler-Smith, 2004:277):

> [h]uman racial classification is of no social value and is positively destructive of social and human relations. Since such racial classification is now seen to be of *virtually no genetic or taxonomic significance* either, *no justification can be offered for its continuance* (Lewontin, 1972, cited in Dawkins, 2004:418 and Edwards, 2003:799, *italics* mine).

As Richard Dawkins puts it:

> [w]e can all happily agree that human racial classification is of no social value and is positively destructive of social and human relations [...] [b]ut that doesn't mean that race is of 'virtually no genetic or taxonomic significance '. [...] However small the racial partition of the total variation may be, if such racial characteristics as there are are highly correlated with other racial characteristics, they are *by definition informative*, and therefore of taxonomic significance (Dawkins, 2004:418, *italics* mine).

Edwards addresses this point in a more technical manner, highlighting that

> [Lewontin's] conclusions are based on the old statistical fallacy of analyzing data on the assumption that it *contains no information beyond that revealed on a locus-by-locus analysis*, and then drawing conclusions solely on the results of such an analysis (Edwards, 2003:799, *italics* mine)

and arguing instead for taking into account the "correlations amongst the different loci, for it is these that may contain the information which enables a stable classification to be uncovered" (Edwards, 2003:799) as offered, for example, by Principal Components Analysis (Cavalli-Sforza *et al.*, 1994:39-42; Tabachnick & Fidell, 2001:582-652) and the classification of matrices of pairwise distances between sequences (Edwards, 2003:799-800).

Edwards concludes that "[t]here is nothing wrong with Lewontin's statistical analyses of variation, only with the belief that it is relevant to classification [...] a proper analysis of human data reveals a substantial amount of information about genetic differences" (Edwards, 2003:801).

There is now a wealth of data showing Edwards to be right and Lewontin to be wrong. For example, Jobling, Hurles & Tyler-Smith (2004), shortly after citing Lewontin's statement on classification, go on and present data supporting the fact that given enough genetic information, the origins of an individual can be determined with a certain probability (Jobling, Hurles & Tyler-Smith, 2004:278-280), but still conclude that "[...] the differences

between them [genetic groups] are too small to justify being *called races*, which would require ≥ *30% difference between groups*" (p. 280, *italics* mine). Where does the 30% threshold come from? And why so much trouble for a *word*? The main take home message is that *there is genetic structure* in living human populations.

Rosenberg *et al.* (2002) report a study of the genetic structure of human population using 377 autosomal microsatellite loci in 1056 individuals from 52 populations (Rosenberg *et al.*, 2002:2381) and found that within-population variation accounts for most of human genetic diversity (93-95%), while inter-population within regions accounts for 3-5% (Rosenberg *et al.*, 2002:2381-2382, especially Table 1:2382). But "[d]espite small among-populations variance components and the rarity of 'private' alleles, analysis of multilocus genotypes allows inference of genetic ancestry without relying on information about sampling locations of individuals" (Rosenberg *et al.*, 2002:2382) and, when applying a model-based clustering algorithm (Rosenberg *et al.*, 2002:2382), for K=5 clusters, they found largely the major geographic regions of the world, and for K=6, they distinguished the isolated Kalash group of Pakistan (Rosenberg *et al.*, 2002:2382). Further on, they tried to find the number of loci required to reproduce the same clustering as when the entire dataset was used, and obtained that, for the Middle East, almost all loci were required, for Oceania and Africa, only ~200, for the Americas, only ~100 while for the entire world sample, only ~150 loci were needed (Rosenberg *et al.*, 2002:2384). "Genetic clusters often correspond closely to predefined regional or population groups or to collections of geographically and linguistically similar population" (Rosenberg *et al.*, 2002:2384), and "[b]ecause most alleles are widespread, genetic differences among human populations derive mainly from *gradations in allele frequencies rather than from distinctive 'diagnostic' genotypes*." (Rosenberg *et al.*, 2002:2384, *italics* mine).

Bamshad *et al.* (2003) analyzed 100 *Alu* insertions[83] and 60 tetranucleotide microsatellites in 206 individuals from sub-Saharan Africa (58), East Asia (67) and Europe (81) and just the 100 *Alu* insertions in a supplementary sample of sub-Saharan and Indian individuals, resulting in a total of 565 individuals from 23 ethnic groups and Indian castes (Bamshad *et al.*, 2003:579). They studied the number of loci required to correctly predict the population of origin of an individual. For entire continents, the mean probability of correct identification

83 The *Alu* insertions are repetitive short interspersed elements (SINEs) with length ~300bp (Jobling, Hurles & Tyler-Smith, 2004:32).

increases rapidly with the number of *Alu* or microsatellite loci used (Bamshad *et al.*, 2003:580). For *Alu* insertions, the mean prediction probability was 40-50% for a single locus and 95-99% for 100 loci, depending on population: "[...] for a given number of loci [*Alu* insertions], it was easier, on average, to distinguish African from non-African than it was to distinguish between Europeans and East Asians" (Bamshad *et al.*, 2003:581). Microsatellites have, on average, the same predictive power as the *Alu* insertions (Bamshad *et al.*, 2003:581), and, combined, they have increased power: with only 160 *Alu* insertions and microsatellites, on average, the correct prediction reached 99-100% for all samples (Bamshad *et al.*, 2003:582). The conclusion is that there is enough genetic structure to allow reliable prediction of population of origin using a limited number of loci. Again, it must be noted that it is not population-specific loci which allow this classification but their correlational structure, confirming Edwards' claims (2003:799).

Another relevant study, from many more not mentioned here, is Long & Kittles (2003); the authors challenge the standard application and interpretation of the F_{ST} statistic to human populations and show that the violation of hidden assumptions results in biases towards reporting increased uniformity.

> F_{ST} measures the extent of subpopulation differentiation as the decrease in heterozygosity relative to that which would be expected if mating were at random throughout the entire population. F_{ST} can be interpreted equivalently as [a] measure of gain of homozygosity (Long & Kittles, 2003:450).

The authors set out to show that the validity of the almost universal finding of inter-population F_{ST}'s of ~15% for the world-wide human population are questionable, because "estimates of F_{ST} will fail dramatically to identify important differentiation among groups, because [...] [it] is strongly biased by violating two hidden assumptions:" (Long & Kittles, 2003:450)

- *the expected gene identity is the same in every population* (Long & Kittles, 2003:450): it is assumed that effective population sizes are equal for all subpopulations, but, if this is violated, then the expected relatedness among alleles will differ across subpopulations (Long & Kittles, 2003:455);
- *the divergence between all pairs of populations is equal and independent* (Long & Kittles, 2003:450): it is assumed that every subpopulation is evolving independently.

> The consequences of variable effective population size and evolutionary nonindependence compound each other. Evolutionary independence cannot be achieved in a hierarchically structured population unless every level is completely

> balanced. For example, each subpopulation must have the same number of individuals and each continent [...] the same number of subpopulation (Long & Kittles, 2003:457).

Moreover, they show that the value of F_{ST} depends on allele frequencies and, thus, is not free to vary from 0.0 to 1.0 (Long & Kittles, 2003:450): F_{ST} can be 1.0 only if every subpopulation is fixed, and F_{ST} can never be very high for genetic loci with high heterozygosity (Long & Kittles, 2003:455). This dependence of F_{ST} on absolute diversity could explain the different results obtained for various genetic markers (Long & Kittles, 2003:466). As an example of the influences of these biases on the estimates of F_{ST}, the authors report that for a large set of dinucleotide repeat polymorphisms, the F_{ST} for humans was 0.119 while for humans and chimps it was 0.183, both not very different from the standard human world-wide 0.15 (Long & Kittles, 2003:450). They conclude that

> [...] the ubiquitous finding $0.10 \leq F_{ST} \leq 0.15$ is due primarily to statistical artifact. There is little meaning to simple partitions of human genetic variation on a world-wide scale, and the broad acceptance of F_{ST} as a valid measure has prevented a deeper understanding of human variation. [...] The patterns of variation within and between groups are too intricate to be reduced to a single summary measure (Long & Kittles, 2003:450, 469).

The overall conclusion, given the results presented above, is that there is genetic structure in human populations, allowing reliable prediction of population membership with a limited number of loci. This information is not encoded in a few population-specific alleles but in the distributional properties of many ubiquitous alleles, making thus Lewontin's claims and following restatements meaningless. What cannot be overstated is that this results *do not validate in any imaginable way racist views or attitudes*. Another important message is that politics sometimes biases and distorts science in complex ways (see for example the arguments in Wolpoff & Caspari, 1997 and Gross & Levitt, 1998; Annex 2).

2.2.8.2. The evolutionary interpretations of modern human genetic diversity

The fact that we are a moderately differentiated species proves to be essential for studies of human evolution, because, if the genetic uniformity would have been (almost) absolute, then no inferences concerning our history could have been made, except that we are *very* recent, indeed.

Usually, and especially in the popularization press, the genetic data on living humans is taken as definitive proof that ROA is true, that there was a single recent speciation taking place in Africa somewhere ~150kya, followed by the spread of modern humans throughout the world with replacement of preexisting archaic humans (e.g., Stringer & McKie, 1996). There are two main types of claims from genetic data taken to support ROA and reject the alternatives: those based on the higher genetic diversity of African populations plus the African roots of modern genetic trees, and those based on the relative genetic uniformity of living humans.

It is generally agreed that most genetic loci are more diverse in sub-Saharan African populations than in all the other populations, and that the non-African diversity is a subset of the African diversity (Jobling, Hurles & Tyler-Smith, 2004:251-252; Relethford, 2001:101-104). Also, for most genetic loci, the reconstructed tree is rooted in Africa and the MRCA is fairly recent (Jobling, Hurles & Tyler-Smith, 2004:251-252; Relethford, 2001:101-104). The standard interpretation is that the modern humans arose recently in Africa and a branch of this population split off and colonized the rest of the world (Stringer, 2002).

But other plausible interpretations of these patterns are available. Greater genetic diversity of the African population can equally well be explained by a longer history of the lineage leading to modern humans in Africa (favored by ROA) or by a long-term greater African population size (Relethford, 2001:137:141; Relethford, 2003:68-70). Also, the African rooting of the gene trees can equally well be accommodated by the African speciation hypothesis or by the greater African population size (Relethford, 2001:137:141; Relethford, 2003:68-70). The stochastic nature of this process is supported by the ancient rooting of the X chromosome markers in Asia (Garrigan *et al.*, 2005b; Ziętkiewicz *et al.*, 2003; Yu *et al.*, 2002; Section 2.2.3). The fact that the long-term human African population was dominant during most of human evolution is supported by palaeoclimatic and ecological models (Relethford, 2001:111-112).

As John Relethford says: "[i]n my views, this finding [small effective long-term population size for humans] is the strongest genetic evidence for replacement" (Relethford, 2001:146). The long-term human effective population size (see Section 2.2.3) is usually evaluated at tens of thousands (Relethford, 2001:146, 151-154; Jobling, Hurles & Tyler-Smith, 2004),

and this figure is considered too low to account for the global gene flow claimed by multiregionalists and taken, thus, to represent definite proof of a bottleneck, usually considered synonymous with a speciation event (Stringer & McKie, 1996; Relethford, 2001; Jobling, Hurles & Tyler-Smith, 2004). Templeton (1998:633) shows that under certain assumptions, interchanging just 1.35 *effective* individuals every generation would still be compatible with a single evolutionary lineage encompassing the entire range inhabited by *Homo* over long periods. But the most plausible account for the apparent disparity between the long-term effective population size and the census population size required to accommodate a world-wide sustained gene flow is provided by *meta-population* models.

"Metapopulations are made up of transient populations connected by migration, subject to extinction and rebirth by colonization, as well as fluctuations in local size" (Harding & McVean, 2004:669) or "[...] metapopulation biology [is that] which concerns itself with the evolutionary effects of a population subdivided into small local populations that frequently become extinct" (Relethford, 2001:171). Such metapopulation models seem to fit quite naturally the pattern of human dynamics over the last 2my, but before the Neolithic, of low population densities, reduced deme size because of the low carrying capacity of the various environments and their high mobility (Relethford, 2001:174-176), inferred from both the archaeological record and living (and historical) hunter-gatherer models. Their main consequences, from a genetic point of view, are that they have "[...] the potential to elevate both the average level of *genomic diversity and expected TMRCA* [time to the most recent common ancestor] estimates, but makes a particularly substantial impact by increasing the *variance in TMRCA estimates*" (Harding & McVean, 2004:670, *italics* mine); and also, as argued by Harding & McVean (2004:670), Relethford (2001:171-176) and Rousset (2003), such metapopulation models could simulate the presence of a bottleneck, as "[t]his process can maintain a large census size, but because of genetic drift introduced by colonization of small numbers from related populations, the overall *effective* size is often quite small" (Relethford, 2001:171, *italics* in original). Such models can naturally accommodate apparently difficult to understand finds, like the large range of TMRCAs for autosomes, the younger NRY's TMRCA compared to mtDNA's and some very low F_{ST} estimates (Harding & McVean, 2004:670).

The comparison between the core population with subsequent radiation and colonization

(incorporated into ROA) with a metapopulation model (as advocated by allotaxa/multiregional approaches) is summarized in the diagram below:

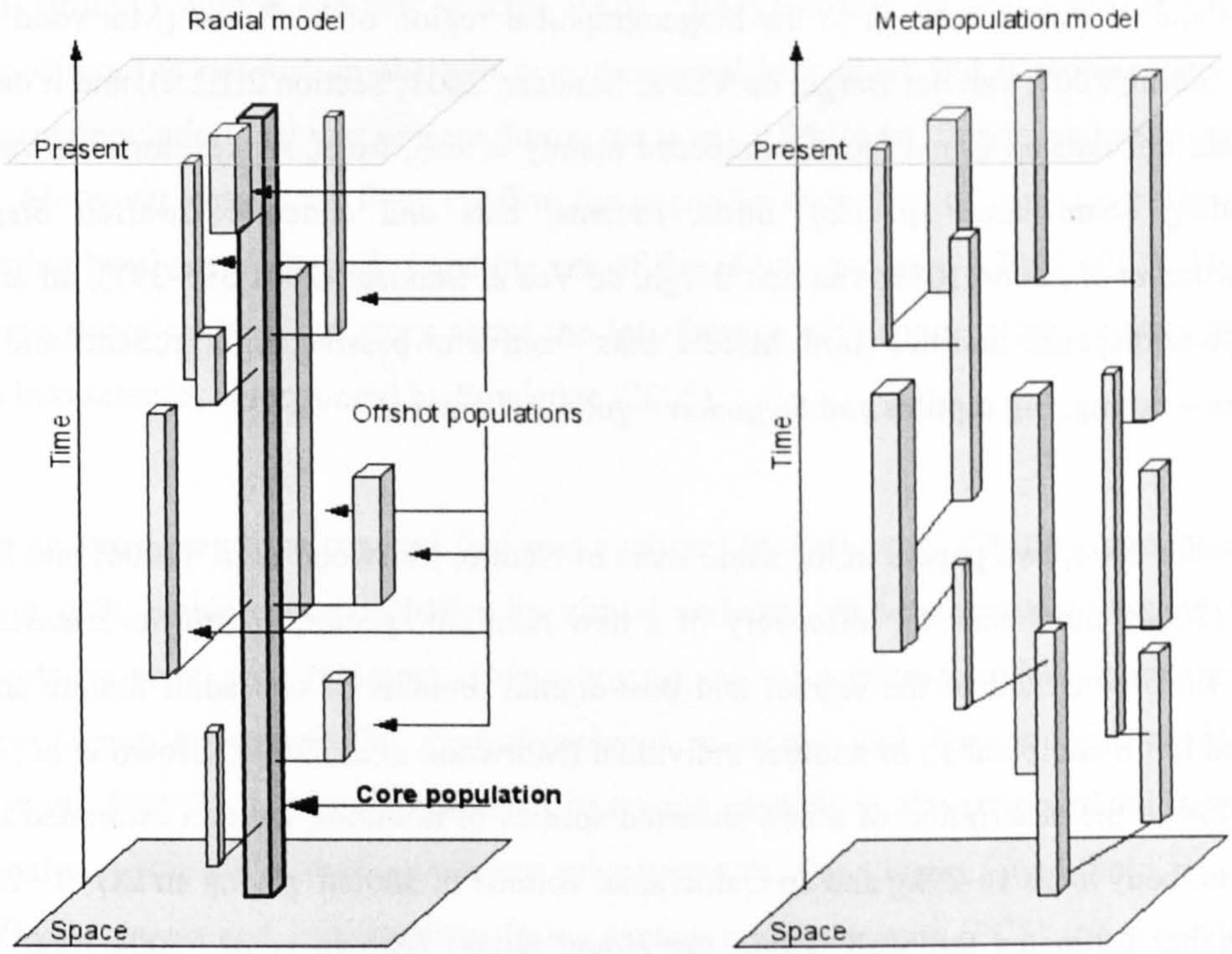

Figure 12: Radial versus metapopulation model.
Adapted from Harding & McVean, 2004:670, Figure 1. Each box represents one population occupying a given location and persisting for a certain period. Box size represents population size. The lines connecting populations represent (re)colonization events.

In the radial model, there is a core population from which offshot populations split and disperse geographically, so that, genetically, this core population is overwhelmingly important, while in a metapopulation model there is no such core population but frequent extinction and recolonization from neighboring (related) populations, so that there is no major player. Translated to the modern human origins problematic, a metapopulation model positing a dominant role for Africa because of its long-term greater carrying capacity, could elegantly account for the current data, both palaeoanthropological and genetic.

It can be concluded, thus, that the genetic diversity of living humans cannot be taken as definitive support for ROA and rejection of its competitors; more than that, it looks like the metapopulation models are a very promising avenue for future research and they seem better able to explain apparent anomalies in the genetic data than the radial model of ROA.

2.2.9. The unexpected diversity of the genus *Homo*: the Flores man

The island of Flores belongs to the biogeographical region of Wallacea (Morwood *et al.*, 1998; Storm, 2001; van der Bergh, de Vos & Sondaar, 2001; Section 2.1.2.2), and it displays a typical unbalanced island fauna, composed mainly of fish, frogs, snakes, tortoise, varanids (including some large species), birds, rodents, bats and endemic dwarfed *Stegodon* (Morwood *et al.*, 2004:1089; van den Bergh, de Vos & Sondaar, 2001:395-397), all animals with good dispersal abilities (bird, insects, bats – active or passive flight, rodents and small reptiles – rafting, big reptiles and *Stegodon* – good swimmers).

In October 2004, two papers in the same issue of Nature, Morwood *et al.* (2004) and Brown *et al.* (2004) announced the discovery of a new hominin species, *Homo floresiensis*. The initial finds consisted of the cranial and post-cranial remains of one adult female and the isolated left mandibular P_3 of another individual (Morwood *et al.*, 2004; Brown *et al.*, 2004) and allowed the description of a new dwarfed species of hominin, with an estimated height ~106cm, body mass 16-29kg and an endocranial volume of 380cm^3 giving an EQ of ~2.4-4.4 (the higher estimates fall well within the *Homo* range) (Brown *et al.*, 2004:1060). This skeleton combines a very small stature and brain size (in the early *Australopithecus* range) with a "unique mosaic of primitive and derived traits in the cranium, mandible and postcranial skeleton" (Brown *et al.*, 2004:1060).

It is proposed that this find represents a true breeding population, and not just some microcephalic individuals, evolved from *Homo erectus* immigrants on the island of Flores through the process of island dwarfing (Brown *et al.*, 2004:1060). The dating of the *Homo floresiensis* sites show that it was present "from before 38 kyr until at least 18 kyr – long after the 55 to 35 kyr time of arrival of *H. sapiens* in the region" (Morwood *et al.*, 2004:1089). The associated stone tools suggest an advanced technology, adapted for big-game hunting (Morwood *et al.*, 2004:1089), including points, perforators, blades and microblades. It is, thus, plausible to suggest that *Homo floresiensis* represents the descendants of earlier *Homo erectus*, attested by finds dated at ~840kya (Morwood *et al.*, 2004:1087), adapted to island conditions and surviving until at least 18kya, overlapping with modern *Homo sapiens*.

The possibility that the finds represent a sample of pathological humans instead of a new hominin species was immediately suggested, but new discoveries reported in Morwood *et al.* (2005), strongly support the new species status: "[they] further demonstrate that LB1 [the type specimen] is not just an aberrant or pathological individual, but is representative of a long-term population that was present during the interval 95-74 to 12 thousand years ago" (p. 1012). Moreover, these new finds confirm the cognitive capacities of *Homo floresiensis*, by suggesting butchery of *Stegodon* and the use of fire (Morwood *et al.*, 2005:1012). The saga of *Homo floresiensis*, a sad story about the interference of human nature, politics and the media into science, is recounted by Powledge (2005).

The brain structure of the original find was analyzed by Falk *et al.* (2005), in an attempt to rule out pathological microcephalia: its virtual endocast was compared to scaled virtual endocasts of great apes (chimps), *Homo erectus*, *Homo sapiens*, modern human pygmy, modern human microcephalic, *Australopithecus africanus* and *Paranthropus aethiopicus* (Falk *et al.*, 2005:242). A new estimate of its cranial capacity is also proposed (417cm^3). In a principal components analysis of various measurements, the authors found that LB1 groups with *Homo erectus* and separate from *Homo sapiens* and the pygmy (PC1) and separate from *Homo erectus* and the microcephalic (PC2) (Falk *et al.*, 2005:243) and after a feature-by-feature comparison, they conclude that *Homo floresiensis* is not a microcephalic modern human, nor a pygmy and that "LB1's well-convoluted brain could not have been a miniaturized version of the brain of either *H. sapiens* or *H. erectus*" (Falk *et al.*, 2005:245). Despite critics (Weber, Czarnetzki & Pusch, 2005; Martin *et al.*, 2006; Jacob *et al.*, 2006), this conclusion seems to stand (Falk *et al.*, 2006; Argue *et al.*, *in press*; Brumm *et al.*, 2006), but it is fair to say that the jury is still out.

The existence of this new species of *Homo* on the island of Flores is not directly relevant to the ROA hypothesis and its competitors, but it does weaken one of its main claims. It seems very improbable that the ancestors of *Homo floresiensis* could have colonized the island of Flores through accidental rafting, mainly because of their large body size and the absence of other large poor swimmers from the fauna of the island. It is, thus, probable that they colonized the island using some form of controlled sea-faring, which would imply that this technology precedes by hundreds of thousands of years the colonization of Australia by modern humans. This technology would probably require high levels of social coordination,

which would, in turn, point to high cognitive abilities and the existence of articulated language. That such capacities can be attributed to *Homo erectus* almost a million years ago, greatly decreases the plausibility of claims of modern human overwhelming superiority (cognitive and/or linguistic) over archaics, allowing a total replacement. Thus, the discovery of *Homo floresiensis* suggests that definitely modern traits, like language, are very old in the human lineage.

2.3. Putting all together: what is the most plausible class of human evolutionary models?

After reviewing the currently most popular human evolutionary model (ROA), its history, its historical competitors and, finally, the main issues surrounding it (either having the potential to falsify it or, against widespread believes to the contrary, unable to support it), I will try to review the class of plausible alternative models and sketch the most probable scenario for human evolution, given the current data. But before proceeding, I must point out that the field is very dynamic, and controversies abound about almost every detail, but, in my opinion, this is a sign of vitality. Over the years, for example, multiregionalism became more refined and continuously updated to reflect recent advances, so that, in its current form (Hawks & Wolpoff, 2001; Thorne & Wolpoff, 2003; Wolpoff & Caspari, 2000; Wolpoff, Hawks & Caspari, 2000; Wolpoff & Caspari, 1997), it acknowledges the central role played by Africa. As Chris Stringer puts it:

> By 1997, Wolpoff and some colleagues had in many respects shifted to a position close to that of the Assimilation Model (Wolpoff & Caspari, 1997). Because this shift was not explicit, I have distinguished it from the original Multiregional Model by the designation 'Multiregional 2' [...]. Multiregional 2 argues that an African influence predominated throughout Pleistocene human evolution because of larger population size, while populations outside Africa were more vulnerable to bottlenecking and extinctions (Stringer, 2002:565).

I must make the point that the shift was not a dramatic one, as implied by Stringer, it even can be regarded as an adjustment of relative weights in a network model; also, it is not implicit (see, for example, Wolpoff & Caspari, 1997:32). Chris Stringer also comments on shifts in his own model: "Some early Recent African Origin [ROA] formulations were implicitly punctuational, with the assumption of a relatively late evolution of a package of 'modern' morphological and behavioural features, and their subsequent rapid spread from

Africa" (Stringer, 2002:564), the shift being also towards allowing "for a greater or lesser extent of hybridization between the migrating population and the indigenous premodern populations" (Stringer, 2002:564).

It would seem, thus, that the two models tend to converge and incorporate new data as they arrive, but it looks to me that the adaptation of Multiregionalism to accommodate these data can be more faithfully described as such, while the incorporation of a moderate to high degree of admixture into ROA strips it of its claims of species status for *Homo sapiens*, thus practically collapsing it into Multiregionalism.

The class of plausible models of human evolution will have to incorporate at least the following main points:

(a) the *allotaxa* status of most[84] of the various "species" of *Homo* during its existence, both in space and time;

(b) pervasive *gene flow* throughout this geographical range[85], allowing:

(c) *synchronized world-wide trends* and

(d) *regional continuity*;

(e) the special role played by *Africa*[86];

(f) *expansion(s) out of Africa* with admixture[87].

There are a number of plausible models proposed in the literature, which I will briefly review below.

2.3.1. John Relethford's *"Mostly Out Of Africa"*

John Relethford (2001:64-65) proposes a two-dimensional classification scheme for human evolutionary models: the *mode of transition* to modern *Homo sapiens*, with two alternatives, multiregional (coalescence in a gene flow network) and speciation, and the *spatio-temporal coordinates* (location and timing) of the transition, also with two alternatives, Africa-recent

84 This does not, in principle, exclude true species status for highly derived variants as, for example, *Homo floresiensis*, but, even in this case, it is still possible that hybridization with modern *Homo sapiens* would have been possible.
85 The actual realization of the potential expressed in (a).
86 Due specially to demographic factors (in turn, effects of ecology/climate).
87 Both demographic expansions and gene spreads have to be considered.

and delocalized (no single, specific place and time for the origin of modern humans). This classification is reproduced in Table 2 below:

		Spatio-temporal location	
		African-recent	***Delocalized***
Mode of transition	***Multiregional***	Primary African Origin Model	Regional Coalescence Model
	Speciation	African Replacement Model (ROA)	impossible

Table 2: Bi-dimensional classification of modern human evolutionary models. *Adapted from Relethford, 2001:64, Figure 3.14).*

This classification manages to distinguish these two very important dimensions, usually amalgamated into a single one, opposing Multiregionalism to ROA (e.g. Lewin, 1998) and also clarifying the claims that Multiregionalism is more of a *framework* than a specific model, allowing (and in need of) refinements and specific sub-models in particular places and at particular times:

> [Multiregionalism] is a *general* explanation for the pattern and process of human evolution within which virtually any hypothesis about dynamics between specific populations can be entertained, from the mixture, even replacement, of some populations to the virtual isolation of others (Wolpoff & Caspari, 1997:32, *italics* in original)

but still remaining fully falsifiable (Wolpoff & Caspari, 1997:33).

The "*Mostly out of Africa*" model of John Relethford (2001:205-211) tries to make sense of the data available so far in this context. He tries to assess the relative probabilities of the various models and to derive his own point of view:

> I feel that the African replacement model [ROA] has a relatively low probability of being correct, in the sense of an origins model in which *Homo sapiens* emerged as a new and reproductively isolated species within the past 200,000 years. [...] However, I agree with those palaeoanthropologists and geneticists advocating a recent African origin of modern *traits* [...]. [...] I suggest that this [the most likely possibility] *is* a multiregional model in the broadest sense of involving genetic input from more than one geographic region within a single evolving species (Relethford, 2001:205-206, *italics* in original).

He goes on and says: "[m]y interpretation of the genetic and fossil evidence is that our ancestry over the past several hundred thousand years is *mostly, but not exclusively out of Africa*." (Relethford, 2001:206, *italics* in original). Below, I reproduce the graphic depiction of his model: "The 'Mostly Out of Africa' model, a multiregional model in which Africa

contributes the most to accumulated ancestry in all regions. [...]" (Relethford, 2001:209, caption of Figure 9.3).

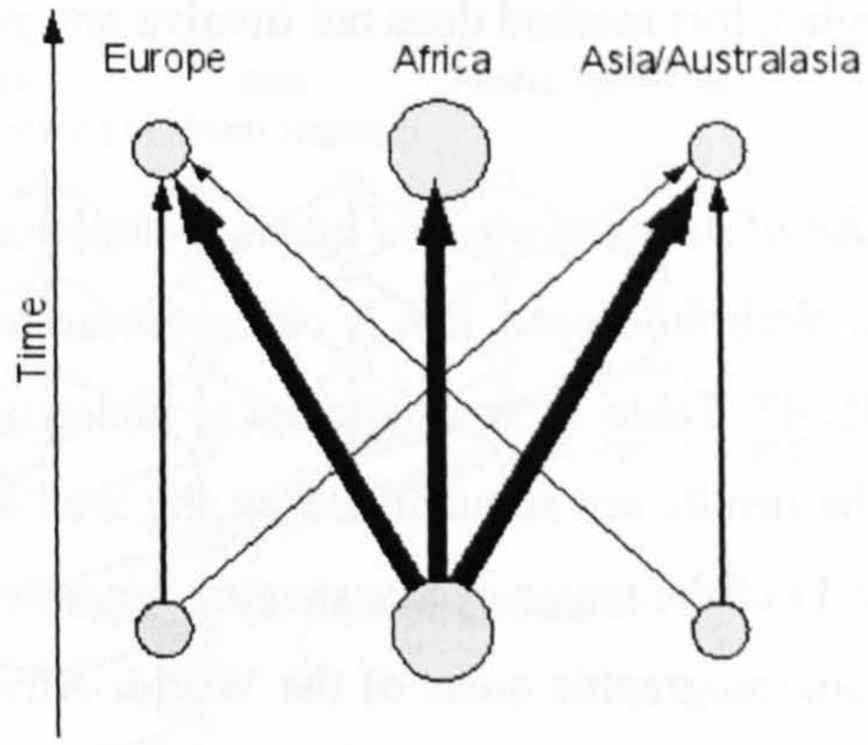

Figure 13: Relethford's (2001) 'Mostly Out of Africa' model.

Adapted from Relethford, 2001:209, Figure 9.3). The size of the circles represents long-term population size and the width of the arrows the relative contributions in terms of accumulated ancestry.

This model was slightly revised in 2003, especially in the light of Templeton's recent work (2002; Section 2.3.2):

> I suggest that 150,000 years ago *most* of our ancestors lived in Africa, but not all of them [...]. I think that the evidence points to *some* ancient non-African ancestry, although it is not clear what was contributed by specific populations from geographic regions outside of Africa (Relethford, 2003:74).

While I basically agree with this proposal (and subtending classification), I think it needs to be made more specific and that, in particular, it lacks back migrations/gene flow into Africa. A much more detailed model, also including backmigrations, is proposed by Alan Templeton.

2.3.2. Alan Templeton's *"Out of Africa again and again"*

Geneticist Alan R. Templeton developed a phylogeographic method that analyzes patterns of genetic diversity and tries to infer the geographical and historical processes having shaped it. The method is called *nested cladistic analysis* and is described in Templeton (1998:642-643) and, briefly, in Jobling, Hurles & Tyler-Smith (2004:193). The method starts from the tips of the phylogeny and incrementally constructs nested clades one mutational step at a time (Templeton, 1998:642). After the entire phylogenetic tree is transformed into a set of nested

clades, those showing significant geographic differentiation are identified and explanations for this pattern are attempted, based on three main processes: gene flow, isolation and expansions (Jobling, Hurles & Tyler-Smith, 2004:193; Templeton, 2002: 45)[88]. As Templeton (2002:45) highlights, this method does not involve any prior model.

In his 2002 *Nature.* paper, *Out of Africa again and again*, Templeton applies nested cladistic analysis to a set of mtDNA, Y chromosome, two X chromosome regions and six autosomal regions (Templeton, 2002:45, 47: Table 3), in an attempt to understand the causes of modern human genetic diversity. The results are summarized by the well-known striking depiction (Templeton 2002:48, Figure 1) of the major events shaping modern human diversity (Figure 14). Represented are the main geographic areas of the World: Africa, Europe (divided into North and South), Asia (divided into North and South), Pacific (including Australia and New Guinea) and the Americas, as vertical lines. Africa's vertical line is thicker, representing its long-term higher population size. The gray arrows stand for major demographic events while the diagonal lines represent gene flow. The time is measured in thousand years before present (kya). The major events are numbered:

(1) the initial out of Africa range expansion of *Homo erectus*, supported by the fossil record;

(2) recurrent gene flow with isolation by distance suggested by *MX1* (chromosome 21) – supported by a single locus;

(3) recurrent gene flow with isolation by distance suggested by Xq13.3, *hemoglobin β*, *ECP* (chromosome 14), *EDN* (chromosome 14), *PDHA1* (X chromosome) – not reaching the 0.05 significance level;

(4) out of Africa expansion suggested by *hemoglobin β*, *MS205* (chromosome 16), *MC1R* (chromosome 16);

(5) recurrent gene flow with isolation by distance suggested by Xq13.3, *hemoglobin β*, *ECP*, *EDN*, *PDHA1*;

(6) out of Africa expansion suggested by mtDNA and the Y chromosome;

(7) out of Asia expansion suggested by the Y chromosome and *hemoglobin β*;

(8) range extensions suggested by mtDNA, *MX1*, *MS205*, *MC1R*, *EDN*;

(9) fragmentation suggested by mtDNA – single locus;

88 The method is not without critics, especially concerning the *inference key* (Jobling, Hurles & Tyler-Smith, 2004:195).

(10)recurrent gene flow with isolation by distance suggested by mtDNA, the Y chromosome, the X chromosome and autosomes.

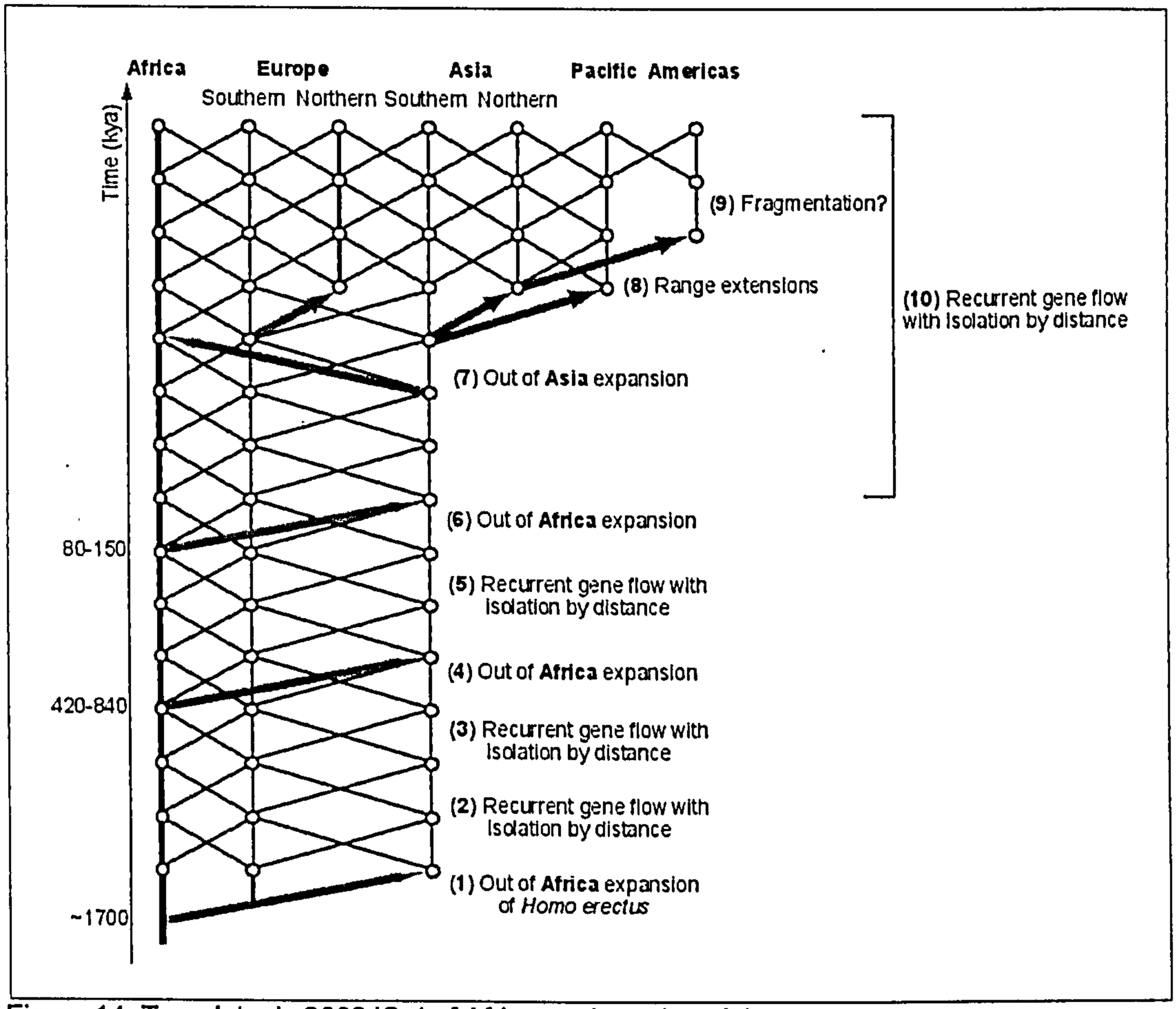

Figure 14: Templeton's 2002 'Out of Africa again and again' .

Adapted from Templeton 2002:48, Figure 1. Detailed explanations are in the main text. Gray arrows represent major demographic events.

It can be seen that there were *many* out-of-Africa expansions[89] as well as expansions back into Africa *from* Asia. The different histories recorded by each genetic marker are integrated into this comprehensive model, suggesting a very complex dynamics of interactions between the main land masses of the Old World, during human evolution.

Alan Templeton summarizes the implications of his model for ROA as "[...] the most recent out-of-Africa expansion event [the one considered by ROA and signaled by mtDNA] was

89 Three are well-supported in this model.

not a replacement event. [...] The hypothesis of a recent out-of-Africa replacement event is therefore strongly rejected" (Templeton, 2002:49). Moreover, he emphasizes that "genetic interchanges among human populations, facilitated both by gene flow and range expansions coupled with interbreeding, has been a major force in shaping the human species and its spatial pattern of genetic diversity" (Templeton, 2002:50). Concerning the "special role that African populations have played in human evolution" (Templeton, 2002:50), he says:

> [t]he genetic impact of Africa upon the entire human species is large because of at least three major expansions out of Africa, although the genetic impact is not as complete as it would be under total replacement. This model is similar to earlier models that have emphasized the role of out-of-Africa population expansion coupled with gene flow and not replacement, such as the assimilation model [...], the multiregional model with expansion followed by admixture [...] and the 'mostly out of Africa' model [...] (p. 50).

There are other models proposed, more or less compatible with the reviewed data, one of them being Vinayak Eswaran's "*Diffusion wave out of Africa*" (Eswaran, 2002). This model proposed that through a shifting-balance mechanism, a coadapted complex of genes has appeared in Africa (Eswaran, 2002:750) and spread over the entire range of *Homo*, because it offered a global selective advantage. This complex represents the modern morphology and its expansion is explained through demic diffusion and admixture (Eswaran, 2002:750), as opposed to population replacement. The proposed global advantage is represented by lowered childbirth mortality (Eswaran, 2002:751), but this mechanism is but a suggestion of the type of selective pressure required. For the moment, I do not have any strong opinion concerning this particular model of human evolution, but I tend to consider it rather improbable, or, at most, as a particular mechanism explaining some special cases. Particularly, I am unconvinced by the nature of this coadapted genetic complex.

2.4. Conclusions

This chapter has reviewed the currently most popular model of modern human evolution, the Recent out of Africa with replacement (ROA), focusing on its history and context of formulation and spread. Its classic alternatives, polygenism and multiregionalism were also presented, each in its own context. Then, a number of issues for ROA were analyzed, ranging from potential refutations of the model to data usually taken to confirm it, but which prove to be unable to distinguish between the competing variants. The profoundly distorting

impact of non-scientific(political and moral) forces on the scientific debate of modern human origins is reviewed in Annex 2, focusing especially on the false and irrelevant accusations of political incorrectness formulated against Multiregionalism, and the issue of racism in human evolution.

I concluded that, against the received opinion, especially in the popularization literature, ROA in its original form is untenable and that it evolved towards the Multiregional framework, by accepting admixture with local archaic populations (thus, denying any speciation claims). Also, Multiregionalism adapted to account for the new data, but in a less dramatic way, the process being more of a fine-tuning nature. Moreover, Multiregionalism must be seen as a general framework accommodating various local scenarios, for each specific case. Relethford's "Mostly out of Africa" (2001) is one such fine tuning, while Alan Templeton's "Out of Africa again and again" (2002) offers a very high level of detail, also in a Multiregional framework.

Probably Templeton's model[90] can also be enriched by a better consideration of the archaeological and fossil record, and also, by using a much larger set of genetic markers. The overall pattern of human evolution seems to be reticulated, inside a single wide-spread polytypic species, composed of small, frequently extinct and replacing demes, allowing for regional continuity and synchronized global trends. It seems, thus, a well supported fact that living humans inherit genes from many different such demes, from many locations in space and time. More research will undoubtedly reveal the detailed histories of each geographic region. But what seems very important is the possibility that not only neutral DNA was inherited into modern populations from local archaic, but also coding genes. This perspective could open new ways towards a proper understanding of human diversity, one of the most valuable features of humanity.

90 This model seems quite well-received and accepted. For example, Dawkins (2004:59-62) uses it for "Eve's Tale", concerning the modern humans.

3. Language-genes correlations

"*Correlation*" means that there is a link between the phenomena or entities of interest, such that they *co-vary in non-random ways* (de Vaus, 2002:267; Howitt & Cramer, 2003:62-79; Tabachnick & Fidell, 2001:53-54). There are different types of language-genes correlations. First, there is the correlation between the *language faculty and the human genetic makeup*, as reflected in the *causal* links between one's genes and one's language faculty. This is studied mainly through behavioural genetic methods (Plomin *et al.*, 2001), involving family aggregation, twin and adoption studies, and focusing on detecting the different genes contributions to variation in the language faculty across individuals. Second, there seem to exist correlations between the *large-scale geographical distribution of languages and human genes* (Cavalli-Sforza, Menozzi & Piazza, 1994), pointing to common mechanisms of spread, differentiation and survival. These correlations are *spurious* in the sense that they are due to one or more deeper causal factors, like geography, demographic processes and/or historical accidents.

These two approaches use different methods applied to different datasets and pursue different goals. One of them studies the genetic causes of variation in the linguistic capacity, uncovering the biological bases of language, treating pathologies and improving our performances, while the other is concerned with understanding our history. But there are also commonalities between them: both require intimate interactions between linguists and geneticists and both study variation.

This chapter will review their core principles, methods and findings and will provide the basis for sketching a unifying approach, promising to shed a new light on the complex interaction between human biology (genes) and culture (language). This connection is provided by the *non-spurious correlations between genes and linguistic features*, whereby differences in the genetic makeup of human populations could influence their linguistic makeup (Chapters 4 and 5).

3.1. The correlations between the capacity for language and the genetic makeup

Almost everybody would agree that there is a fundamental correlation between human genetics and the linguistic faculty (Pinker, 1995): language is a species-specific phenotype, found only in humans and displayed by all normal individuals. Thus, there must be something in our genes making us capable of acquiring and using language. Of course, this does not imply in any way a direct hard-coding of language in our genome. This hypothesis is easily ruled out by a huge amount of facts concerning language acquisition, linguistic diversity and language change (Pinker, 1995). But what is then the complex relationship between the genes, the environment and the language faculty? How can we start to disentangle their relative contributions? And what do we know so far?

Behavioral Genetics is the "specialty that applies [...] genetic research strategies to the study of behavior, such as psychiatric genetics (the genetics of mental illness) and psychopharmacogenetics (the genetic of behavioral responses to drugs)." (Plomin *et al.*, 2001:xvii). It can be considered a branch of genetics applied to behavioral sciences, using specific methods to disentangle the roles of genes and environment on the behavior. It has many commonalities with Quantitative Genetics, which concerns itself with the genetics of quantitative traits ("traits that show a continuous range of phenotypes" - Halliburton, 2004:525), because behavior is rarely determined by a few genes[91]. Fundamentally, behavior genetics studies the relationship between *variation* in genes and *variation* in behavior *across* individuals.

3.1.1. Methods

The main idea is that if the variation in genes explains some of the variation in behavior, then one can study the variation in behavioral traits when controlling for the degree of genetic and environmental variation. One important note is necessary at this moment: it is known that our species is genetically very homogeneous (Section 2.2.8); then, what meaning can be attributed to the study of the relationship between genetic and behavioral variation?

91 There are certain *single-gene diseases*, like Huntington's disease, for which the inheritance pattern is Mendelian (Plomin *et al.*, 2001:61).

Humans are not genetic clones, and there are differences between any two individuals, excluding the special case of true twins. There is a huge number of polymorphic loci in humans[92] (Jobling, Hurles & Tyler-Smith, 2004:46-118), but it is currently unknown how many of them are silent. Nevertheless, for any two individuals chosen at random, there is a baseline probability of sharing the same allele at a given polymorphic locus, which tends to increase with the increasing degree of relatedness, as sharing a recent common ancestor increases the chances of also sharing its alleles (Jobling, Hurles & Tyler-Smith, 2004; Chapter 2). Thus, close relatives are more "genetically alike" than distant relatives, which are more "genetically alike" than random individuals. This intuition was used, for example, by Hamilton (1964) to explain the emergence and maintenance of *kin-based altruism*, formulating the theory of *inclusive fitness*[93] (Skelton, 1993:251-252; Dawkins, 1990b:90-108).

The *coefficient of relatedness* represents the probability that two individuals share an allele through common descent (Skelton, 1993:234-237). For any autosomal locus, the child inherits one allele from the mother and one from the father, so that any parental allele has a 50% chance of being transmitted to this child. For a locus on the sex chromosomes, the inheritance pattern is different, because the Y chromosome always comes from the father while one X chromosome is maternal and, for females, the other is paternal. A locus on the mtDNA always comes from the mother (Chapter 2; Jobling, Hurles & Tyler-Smith, 2004:33; Plomin *et al.*, 2001:6-60). Skelton (1993:235: Figure 6.1) lists the coefficients of relatedness for various degrees of kinship.

The *fraternal* or *dizygotic twins* (DZ) develop from two separate ova, fertilized by two different spermatozoa, and are genetically like just two normal siblings: their relatedness being 50% (also 50% chance of having the same sex). The *identical* or *monozygotic twins* (MZ) develop from the same fertilized ovum, following a division of the zygote into two embryos, and have identical genomes (except for *de novo* mutations): their relatedness is

92 For example, the NCBI dbSNP database (Build 125, 7 March 2006) lists 27,189,291 SNP submissions for *Homo sapiens* (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi) while ALFRED lists 1479 polymorphisms at 683 loci (http://alfred.med.yale.edu/alfred/alfredsummary.asp), some of them common between the two databases.

93 *Inclusive fitness* represents an extension of the classic fitness of an individual to also include contributions from the related individuals (Hamilton, 1964; Dawkins, 1990b; Pinker, 1997:398-401).

100% and they have the same sex (except the very rare cases of an XXY zygote splitting into XX and XY embryos). Depending on the exact timing of the split, the twins can share the same *amnion*[94] (monoamniotic) or not (diamniotic), and the same *placenta*[95] (monochorionic) or not (dichorionic). Monochorionic twins experience the same prenatal environment, including nutrient availability, toxic substances and infections, increasing their likeness, on one hand, while competing for the same limited placental resources, decreasing it, on the other (Bishop, 2003:S145; Stromswold, 2001:688).

There are a number of methods designed to allow controlled studies of the interaction between genes, environment and behavior, but some of them are not applicable to humans (*genetic engineering* techniques, such as *gene knockout*[96], *cloning*[97] or *selective breeding*[98]) for ethical reasons. The methods which can be used to study human behavioral genetics, can be classified as follows (Plomin *et al.*, 2001:72-84; Felsenfeld, 2002:334; Bishop, 2003:S144-S147; Stromswold, 2001):

Adoption studies: probably the best-known behavioral genetics methodology, whereby the behavioral correlations between the adoptees and their foster families are compared to the corresponding correlations between the adoptees and their biological family (Plomin *et al.*, 2001:72-75; Bishop, 2003:S145-S146). The genetic relationship between the adoptees and their foster family should reflect the expected relationship between random individuals, while the relatedness with their biological family (parents and sibling) is 50%. By comparing the adoptees' behaviour with that of the foster and biological families, the influence of the shared genes versus the influence of the shared environment can be detected (Plomin *et al.*, 2001:72-75), but several arguments have been adduced against simplistic interpretations of adoption studies, including the effect of prenatal and early postnatal shared environments (Plomin *et al.*, 2001; Stromswold, 2001).

94 A layer of cells surrounding and protecting the embryo (Seeley, Stephens & Tate, 2005:1087).
95 A temporary organ specialized in the nutrient and waste exchange between the embryo and the mother (Seeley, Stephens & Tate, 2005:1085).
96 A specific gene is inactivated and the effects are studied *in vivo* (Hertzog & Kola, 2001:1).
97 Two or more genetically identical individuals are created; a specific technique is the *inbred strains* (Plomin *et al.*, 2001:68-72).
98 Certain phenotypes are allowed to differentially breed, creating an artificial directional selective pressure (Plomin *et al.*, 2001:65-68).

Family aggregation studies: if a trait is influenced by shared genes, then it should "run in families" (Stromswold, 2001:650; Plomin *et al.*, 2001; Fisher, Lai & Monaco, 2003:58). A *proband* is an individual manifesting the considered trait (Stromswold, 2001); thus, such a trait should have a greater incidence among the relatives of a proband as opposed to the relatives of a control (an individual not displaying the trait under consideration; Stromswold, 2001:650):

> more probands than controls should have a positive family history for [trait manifestation] (i.e. more probands should have at least one [trait-manifesting] relative) and a higher percentage of probands' relatives than controls' relatives should have a history of [trait manifestation] (Stromswold, 2001:650).

Well-known examples of such cases are represented by single-gene pathologies like Huntington's disease (autosomally inherited disease – locus 4p16.3, increased length of CAG triplet repeat – characterized by progressive and selective neuronal death with dementia and loss of motor control; OMIM 14340; Plomin *et al.*, 2001:6), phenylketonuria (genetic deficiency of phenylalanine metabolism – locus 12q24.1; PKU; OMIM 261600; Plomin *et al.*, 2001:7) and the pathology identified in the KE British family (OMIM 602081), caused by a mutation in the *FOXP2* gene (locus 7q31). One of the caveats is that not only are genes shared by families, but also the environment.

Pedigree studies can be considered as a refinement of the familial aggregation studies (Bishop, 2003:S146; Stromswold, 2001:696), whereby the actual pattern of transmission is analyzed in affected families across generations and individuals. If the pattern is simple enough, hypotheses can be formulated concerning the inheritance mechanism. For example, in the case of the KE family, the inheritance mechanism is compatible with a single dominant autosomal gene (Bishop, 2003:S146). The difficulties concern how simple the pattern of manifestation of the trait actually is and the amount of available information concerning the family members.

Twin studies: the most powerful methodology for studying the effects of genetic and environmental variation on behaviour (Plomin *et al.*, 2001:75-82; Bishop, 2003:S144; Stromswold, 2001; Plomin & Kovas, 2005). MZ twins share 100% of their genes while DZ twins share only 50%, on average, while the environmental differences experienced by MZ and DZ twins are minimal (Plomin *et al.*, 2001:82), which is not the case when comparing MZ twins with non-twin siblings of different ages, for example. Moreover, by controlling for

same-sex DZ twins, this other important source of variation is equalized. If a trait is under genetic control, then it would be expected that the MZ twins are more alike than DZ twins and, roughly, the amount of difference between the MZ correlation versus DZ correlation corresponds to how strong the genetic influence is. There is an impressive number of twin studies conducted to date, targeting traits covering learning disabilities, written and spoken language, general cognitive abilities or psychiatric disorders (Plomin *et al.*, 2001; Stromswold, 2001; Plomin & Kovas, 2005; Bishop, 2003; Felsenfeld, 2002; Inoue & Lupski, 2003; Plomin, Colledge & Dale, 2002).

3.1.2. Measuring the effect of genes and environment: the *heritability*

The methods presented above concur in suggesting that there are genetic influences on many behavioral and cognitive traits, including language (Stromswold, 2001; Plomin *et al.*, 2001; Bishop, 2003). But "it is possible to ask not only *whether* genetic influence is important but also *how much* genetics contributed to the trait" (Plomin *et al.*, 2001:85, *italics* in original). This involves measuring how much of the inter-individual phenotypic variation for the considered trait in a population can be accounted for by inter-individual genetic variation in the same population (Plomin *et al.*, 2001:85; Stromswold, 2001:652). This refers to the inter-individual differences in the population and not to specific individuals: for example (Plomin *et al.*, 2001:85), untreated phenylketonuria has a devastating effect on the homozygous carriers, including on their cognition (Plomin *et al.*, 2001:85, 6-7, 9-14; OMIM 261600), but due to very low frequency[99], its overall effect on the variation in cognitive abilities in a population is very small (Plomin *et al.*, 2001:85).

Heritability is defined as *the proportion of the phenotypic variation accounted for by genetic variation* (Plomin *et al.*, 2001:85; Stromswold, 2001:652; Halliburton, 2004:539)[100].

A genetic locus L can have one or more alleles, A_1, ..., A_n. L is said to be *associated with a trait* if some of its alleles are associated with different means of the trait in a population (Plomin *et al.*, 2001:343). An individual's phenotype generally results from the complex interaction of the genetic and environmental effects (Plomin *et al.*, 2001:345-346;

99 Plomin *et al.* (2001:85) give a figure of ~0.01%, but PKU varies widely across populations (OMIM 261600).

100The following is based on Halliburton (2004:525-587) and Plomin *et al.* (2001:343-351).

Halliburton, 2004:531-533; West-Eberhard, 2003), which we can symbolize as (Plomin *et al.*, 2001:345; Halliburton, 2004:531):

$$P = G + E$$

where P is the phenotypic trait, G is the genetic effects and E is the environmental effects. The genetic effects can be further divided (Plomin *et al.*, 2001:345; Halliburton, 2004:533):

$$G = A + D + I$$

where A represents *additive genetic effects*, D represents the *dominance effects* due to interaction between alleles at the same locus[101] and I represents the *epistatic effects* due to interaction between alleles at different loci[102]. The additive effect is related to the average effects of single alleles (Plomin *et al.*, 2001:343): it reflects the sum of the average effects of each allele separately (Plomin *et al.*, 2001:343; Halliburton, 2004:535).

Let's consider that L has only two alleles, A_1 and A_2, with the following phenotypic values for the three possible genotypes at this locus:

- A_1A_1 (homozygous for A_1) → phenotype value a;
- A_2A_2 (homozygous for A_2) → phenotype value $-a$;
- A_1A_2 (heterozygous) → phenotype value d (dominance)[103].

The frequencies are p for A_1 and $q = (1 - p)$ for A_2, and, making the assumption that the population is in Hardy-Weinberg equilibrium[104], the frequencies of the three possible genotypes are: $A_1A_1 : p^2$, $A_2A_2 : q^2$ and $A_1A_2 : 2pq$. The mean value of the trait is:

$$\mu = p^2a + 2pqd + q^2(-a) = a(p - q) + 2pqd$$

A genotype's *relative phenotypic value* is its deviation from the population's mean:

$$G_{11} = a - \mu = 2q(a - pd)$$

$$G_{22} = -a - \mu = -2p(a + qd)$$

$$G_{12} = d - \mu = a(q - p) + d(1 - 2pq)$$

These are called *genotypic values*, they measure the genotypes relative to the population

101For all diploid loci there are two alleles in any individual, interacting in various ways (recessive, dominant, etc.).

102Another factor could be added, accounting for *maternal and paternal effects*, but this can be generally neglected.

103Any triplet of values (α, β, δ) can be translated into (-a, a, d), with a = α-(α+β)/2 = (α-β)/2, [-a = (β-α)/2 = β-(α+β)/2] and d = δ-(α+β)/2.

104An abstract and simplified model of a population; some assumptions are: no natural selection, negligible mutation rates, infinite panmictic population. This is a good approximation of some real cases (Halliburton, 2004:69-88).

means, and *depend on allele frequencies* (Halliburton, 2004:534). The average value of A_1 across all the genomes it can appear is:

$$\mu_1 = pa + qd$$

as it can combine with probability p with an A_1 producing an A_1A_1 genotype of phenotypic value a, and with probability q with an A_2 producing an A_1A_2 genotype of phenotypic value d. For A_2 this is:

$$\mu_2 = pd - qa$$

The deviation of each allele from the population mean, μ, is called the *average effect* of the allele (Halliburton, 2004:535):

$$\alpha_1 = \mu_1 - \mu = q(a + d(q - p)) = q\alpha$$

$$\alpha_2 = \mu_2 - \mu = -p(a + d(q - p)) = -p\alpha$$

where

$$\alpha = a + d(q - p)$$

The *additive effect of a genotype* is defined as the sum of the average effects of the two alleles composing it (Halliburton, 2004:535):

$$A_{11} = \alpha_1 + \alpha_1 = 2q\alpha$$

$$A_{22} = \alpha_2 + \alpha_2 = -2p\alpha$$

$$A_{12} = \alpha_1 + \alpha_2 = \alpha(p - q)$$

The additive effects and the genotypic values are different if there are any dominance effects ($d \neq 0$). In general:

$$G_{11} = A_{11} + D_{11}, \text{ with } D_{11} = -2q^2d$$

$$G_{22} = A_{22} + D_{22}, \text{ with } D_{22} = -2p^2d$$

$$G_{12} = A_{12} + D_{12}, \text{ with } D_{12} = 2pqd$$

where D_{11}, D_{22} and D_{12} represent the *dominance effects* (Halliburton, 2004:536). This represents the derivation of the split of genetic effects into additive and dominance components:

$$G = A + D$$

The additive component refers to the average independent contribution of each allele to the phenotypic values, while the dominance effects are related to the interactions between them. It is very important to highlight again the fact that *these effects are considered against the background of the population mean*, and are *crucially dependent on the allele frequencies*:

thus, in a different population with different allele frequencies, the estimations of G, A and D will be different.

To address the contribution of genetic variability to phenotypic variability, the variance of these measures must be considered (the *variance components approach*, Plomin *et al.*, 2001:346). Let us denote by $Var(X)$ the variance of a random variable X (de Vaus, 2002:224; Halliburton, 2004:593-612); then the phenotypic variance can be decomposed as (Halliburton, 2004:607):

$$Var(P) = Var(G + E) = Var(G) + Var(E) + 2Cov(G, E)$$

If only the additive and dominance effects (neglecting epistatic effects) are considered:

$$Var(G) = Var(A + D) = Var(A) + Var(D) + 2Cov(A, D) = Var(A) + Var(D)$$

as, by definition, the additive genetic effects are independent of the dominance effects, $Cov(A, D) = 0$ (Plomin *et al.*, 2001:346; Halliburton, 2004:538, Box 13.1). $Var(A)$ is called the *additive genetic variance* and $Var(D)$, the *dominance variance* (Halliburton, 2004:539). Concerning the *covariance between two individuals*, it can be shown that the genetic covariance between relatives is given by the following table:

Familial relationship	*Symbol*	*Proportion of additive genetic variance, Var(A), shared*	*Proportion of dominance genetic variance, Var(D), shared*	*Symbolic representation Cov(X,Y) =*
Parent-offspring	**PO**	**0.5**	**0**	**$0.5Var(A)$**
Full siblings	FS	0.5	0.25	$0.5Var(A) + 0.25Var(D)$
Half sibling	HS	0.25	0	$0.25Var(A)$
Grandparent-grandchildren	GG	0.25	0	$0.25Var(A)$
First cousin	FC	0.125	0	$0.125Var(A)$
Non-identical twins	**DZ**	**0.5**	**0.25**	**$0.5Var(A) + 0.25Var(D)$**
Identical twins	**MZ**	**1**	**1**	**$Var(A) + Var(D)$**

Table 3: Covariances (coefficients of relatedness) among relatives, expressed as function of the additive and dominance genetic variances.
Adapted from Halliburton (2004:540-542, Table 13.5) and Plomin et al. *(2001:348-349, Table A.1). In* **bold** *are the most important types of relationships for behavioral genetic research.*

But the members of a biological family share more than just genes: the shared environmental factors which tend to make the family members more alike for the trait under consideration

are called *shared environmental influences*, while the *nonshared environmental influences* do not result in the members becoming more similar (Plomin *et al.*, 2001:348-349). Most behavioral genetic studies focus on the additive genetic variance, shared and nonshared environmental influences; this approach is known as the *ACE model* (Additive genetic effects + Common (shared) environment + nonshared Environment) (Plomin *et al.*, 2001: 349). Examples of shared environmental influences include the linguistic input from parents (Stromswold, 2001:652) or nutrition and hygiene habits, while nonshared environmental influences include events and/or processes peculiar to a single individual, like accidents, illnesses (Stromswold, 2001:652) or separate peer groups.

The *broad sense heritability* (Halliburton, 2004:539; Plomin *et al.*, 2001:349) is defined as the ratio of the total genetic variation (additive or not) to the phenotypic variation:

$$H^2 = Var(G) / Var(P)$$

while the *narrow sense heritability* is the ratio of just the additive genetic variation to the phenotypic variation:

$$h^2 = Var(A) / Var(P)$$

$H^2 > h^2$, for $Var(D) \neq 0$ and $Var(I) \neq 0$, otherwise $H^2 = h^2$. Usually, *heritability* is understood as narrow sense heritability, unless otherwise specified (Halliburton, 2004:539), and reflects the extent to which a trait will be transmitted from parent to offspring (the expected degree of their similarity on genetic grounds). The broad sense heritability represents the influence of any kind of genetic factors on the trait variation in the population (Plomin *et al.*, 2001:349). Halliburton (2004:540, Table 13.4) lists a series of heritabilities for various traits in different species, while Plomin *et al.* (2001) discuss heritabilities of cognitive, psychological and psychiatric traits in humans.

Estimating heritabilities using the various methodologies listed in Section 3.1.1 is complex and will be illustrated with twin studies, by far the most used method[105] (Plomin *et al.*, 2001:351). The idea is to compare the concordance rates or correlations[106] between MZ and DZ twins (Stromswold, 2001:652; Plomin *et al.*, 2001:351) in the population, for the

105For a parent-offspring example and subtending assumptions, see Halliburton (2004:543).

106*Concordance rates* are used for binary traits (i.e., present/absent, diagnosed with a pathology or not) and represent the frequency with which the co-twins have the same value for the trait (i.e., both absent) across the population. *Correlations* are used for continuous traits (i.e., height, vocabulary size) and measure (usually) Pearson's *r* between co-twins across the population (Stromswold, 2001:652-655).

considered trait. It can be shown (Plomin *et al.*, 2001:349-351) that the concordance rate/correlation between MZ twins, r_{MZ} and between DZ twins, r_{DZ}, can be expressed as:

$$r_{MZ} = h^2 + c^2$$

$$r_{DZ} = 0.5h^2 + c^2$$

where c^2 represents the shared environmental influences (Plomin *et al.*, 2001:349-350), so that:

$$h^2 = 2(r_{MZ} - r_{DZ}).$$

This measure of the narrow sense heritability is called *Falconer's estimate of heritability* (Stromswold, 2001:655). Moreover,

$$e^2 = 1 - r_{MZ}$$

where e^2 represents the nonshared environment; given that the MZ twins are genetically identical, any variance not explained by the shared environmental factors *must* be attributable to the nonshared environment (Plomin *et al.*, 2001:350; Stromswold, 2001:655). For the complex process of estimating heritabilities in real cases, see Plomin *et al.* (2001:351-368) and Stromswold (2001).

First, let us consider a *trait which shows no variation in the population*, all individuals having the same value for this trait. A good example could be the number of hearts: all individuals in the population have one and only one. Applying the definition of heritability (narrow or broad), we obtain that it is undefined, as the denominator, *Var*(*P*), is 0. But on independent grounds, we can certainly consider the number of hearts as genetically determined[107]. Thus, *there are genetically determined traits for which it is meaningless to talk about heritability*. Conversely, consider the trait defined as the capacity of teleportation to Mars and back: it is clear that the trait is uniformly absent in human populations, as there is no reliable report of people being able to do so. Again, phenotypic variability is 0 in population and the heritability of the trait in undefined. But, in this case, we intuitively know that this trait is not genetically encoded but a non-genetic consequence of the physical laws governing complex macromolecular systems. Thus, *there are also non-genetically determined traits for which it is meaningless to talk about heritability*. Concluding, the concept of heritability is meaningless for uniform traits and cannot be used to infer anything about the genetic basis of such a trait. More importantly, this highlights the fact that

107This also involves a judgment based on variation, but this time across taxa (orders, kingdoms) and evolutionary time.

heritable and genetic are different concepts.

Second, *the heritability of a trait in the case of no environmental variation is higher than for the same trait in the case of a variable environment.* This counter-intuitive consequence can be exemplified (following Bishop, 2003:325) by considering a thought experiment involving height. Suppose that a random half of the population receives a systematic dose of growth hormone, while the other half does not. We know that in real populations height is highly heritable (Bishop, 2003:325; Halliburton, 2004:540, Table 13.4), but during this experiment, height would appear as unaffected by genes simply because of the very strong impact of the environmental differences (growth hormone administration versus non-administration). In a second phase, everybody gets the growth hormone administered: everybody's height increases but the heritability is high again as the environmental effect, even if very strong, is now uniform in the population. Another example involves muscle strength/mass and exercise: if nobody does systematic exercises, the muscle strength/mass is highly heritable. If some of the people exercise regularly and some don't, then the heritability is low, because of the overwhelming effect of these environmental differences, while if everybody has the same regime of physical exercising, heritability goes high again, because the environmental effect is the same for everybody. As Bishop (2003:325) puts it, these are not simply fictitious examples, as the real average gain in height in Western (and westernized) societies during the last decades proves: improved health, hygiene and nutrition only for some have decreased heritability estimates, while their generalization restored high heritabilities (but with an increased average). What these cases show is that *heritability is not an absolute measure of some kind of intrinsic genetic contribution to the trait but a relative estimate crucially dependent on the specific population and environmental conditions.* The entire issue of some kind of intrinsic genetic contribution to a trait is simply misleading and unreal, as shown extensively by phenotypic plasticity (West-Eberhard, 2003; Gerhart & Kirschner, 1997), developmental biology (Gilbert, 2000), psychology (Plomin *et al.*, 2001) and many other disciplines. There can be no phenotype without both genes and environment and even the distinction between genetic information and environment is fuzzy in many cases (West-Eberhard, 2003; Gerhart & Kirschner, 1997).

Plomin *et al.* (2001:87-88) offer an intuitive illustration of the fact that the interaction between genes and environment is intrinsic to the phenotype but, still, that we can talk about

genetic and environmental contribution in a meaningful way. Consider a rectangle of length *l* and width *w*; it has by definition an area $a = l * w$. We can equate its area *a* with the measure of the phenotypic trait, the length *l* with the genetic contribution and the width *w* with the environmental contribution, respectively (Figure 15). It is meaningless to talk about the separate contributions of length and width to area in any individual case (i.) (i.e., for any particular rectangle) as there can be no area without both length and width. However, if we consider instead a (non-uniform) population of rectangles then the variation in area can be due to:

ii. variation only in length: *w* is constant, *l* is variable;

iii. variation only in width: *l* is constant, *w* is variable;

iv. variation in both: *l* and *w* are variable.

The analogy is useful as it is also relevant to the case discussed previously: if the population is phenotypically uniform (all the rectangles have the same area), this can be due to:

v. uniform length and width: *l* and *w* are constant;

vi. compensation between *l* and *w*: both *l* and *w* are variable but they vary so as their product remains constant, $l * w$ is constant.

Case (v.) is uninteresting but case (vi.) illustrates the very important fact that *even when heritability is high and there is an important contribution of genes to a trait, environmental manipulations can make this irrelevant*. More specifically, even if some pathology (e.g. PKU) has a very strong genetic component, environmental approaches can successfully address it. The same is also true for cognitive or psychiatric pathologies, where appropriate environmental interventions can dramatically increase the performances and quality of life of the affected individuals (Plomin *et al.*, 2001; Stromswold, 2001; Inoue & Lupski, 2003).

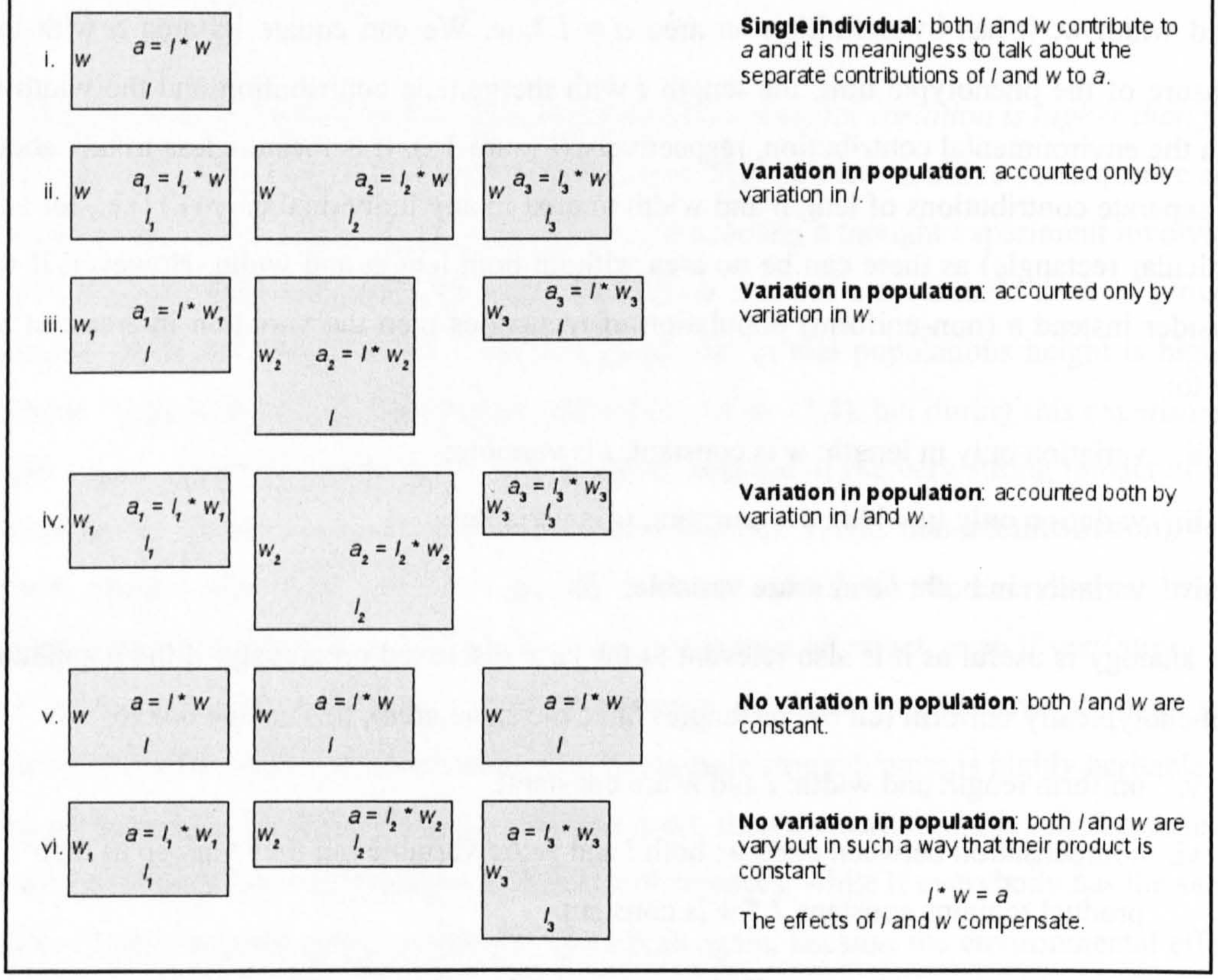

Figure 15: Illustrating the genetic and environmental effects on the phenotype

A rectangle represents one individual, its area a *is the measure of the trait of interest, the length* l *represents the genetic effects and the width* w *represents the environmental effect. See text for details. Adapted from Plomin et al., 2001:88, Figure 5.13.*

Third, suppose a *very important gene*, which is uniform in population exactly because of its importance (stabilizing selection), is mutated: its effects on the phenotype are devastating, usually lethal, even if it does not normally show up in any heritability measures for the affected traits. An example at hand is represented by the disruption of *FOXP2*.

Fourth, returning to the hypothetical example of the growth hormone administration and the real case of increase in height during the last decades, even if inside each group (growth hormone administered vs not administered and improved nutrition, health and hygiene vs bad nutrition, health and hygiene) the heritability of the trait is high, *the differences between groups (in average height) is not accounted by genes but purely by environmental variation.* That is, equalizing the environmental factor totally eliminates the inter-group differences.

This is a very important point to consider when comparing heritabilities across populations[108], as "[t]he causes of average differences between groups are not necessarily related to the causes of individual differences within groups" (Plomin *et al.*, 2001:89), so that finding high heritability for normal individuals for a trait does not necessarily imply high heritability for the pathologic forms of that trait (Plomin *et al.*, 2001:89), but the converse is also true: it might turn out that differences in heritabilities between two groups are due to differences in environmental conditions, but identical genetic factors.

Fifth, *heritability can change with age*. For example, during the early and late stages of development, different process are involved, subtended by different genetic factors. Or, alternatively, in various stages the environmental variation might differ for the specific trait, let's say, more uniform in the early stages than in the later. Another process could be represented by a positive feedback effect between genetic tendencies and active environmental search and construction. A well-known example of change in heritability is provided by the dramatic increase in the heritability estimates of *g* with age (Plomin *et al.*, 2001:173-177; Jensen, 1998). During childhood, the shared environment is very important, but it declines with increasing age, so that for adults, the genetic factors and the nonshared environment account for most of the variation in the general cognitive abilities (Plomin *et al.*, 2001:177)[109].

The following section will discuss heritability estimates for language and speech, but a preliminary technical note is necessary in order to proceed. Some extreme values of Falconer's heritability estimates (h^2), below zero or above one, may result if there are interactions between genetic and environmental factors. If MZ twins are treated more similarly than DZ twins, the heritability estimates may be greater than one, while if MZ twins compete for resources more than DZ twins, the heritability may turn out negative[110] (Stromswold, 2001:660, Footnote 10).

108And also when dealing with simplistic and racist accounts of the "genetic differences between races", IQ being a predilect case.

109Which could be (at least, partially) explained by the same type of active environmental construction and positive feedback between environment and genes.

110Due to limited resources trade-offs (e.g., nutrients), which makes them more dissimilar than non-competing DZ.

3.1.3. Heritability estimates for speech and language

There are a number of studies addressing the heritability of different aspects of speech and language but the best review to date remains Stromswold (2001) (Wagner, 2002, offers a short overview). Karin Stromswold, after reviewing more than 100 studies involving the genetics of language and speech, concludes that:

> [...] genetic factors account for some of the individual differences in linguistic ability for both normal people and people who suffer from developmental language disorders. [...] heritable factors typically accounted for over half of the variance in language-impaired people's linguistic abilities. [...] [normal] MZ twins are linguistically more similar to one another than DZ twins for all aspects of written and spoken language. [...] specific-to-language genetic factors play a substantial role in the variation observed in linguistic abilities among both people who suffer from language disorders and those who do not (Stromswold, 2001:704-705),

conclusions which are supported by later studies. I will now briefly review the main findings concerning the heritable factors in language and speech[111].

The best studied aspects concern speech and language pathologies (Stromswold, 2001; Bonneau, Verny & Uzé, 2004; Bishop, 2003; Fisher, Lai & Monaco, 2003; Felsenfeld, 2002; Plomin, Colledge & Dale, 2002; Plomin & Kovas, 2005). Plomin, Colledge & Dale (2002:419-420) report concordances of ~0.85 for MZ and ~0.50 for DZ (giving $h^2 \approx 0.70$) for language generally, from the literature. They also report the findings of the TEDS study[112] that on a composite measure of language disability, MZ concordances are 0.88 and DZ concordances are 0.51 (giving $h^2 \approx 0.74$).

Felsenfeld (2002:335) reports that 70% of the variance in liability to stuttering can be accounted for by additive genetic effects ($h^2 \approx 0.70$). Plomin & Kovas (2005:595, Table 1, 596-597, Table 2) report heritability estimates from the literature for various aspects of language disabilities. For auditory repetition impairments they report $h^2 \approx 0.12$, $h^2 \approx 1.18$ for non-word repetition, $h^2 \approx 0.26$, 0.34 and 0.38 for different composites of language tests at different ages. Concerning reading disabilities, they report $h^2 \approx 0.60$ for Neale reading test, h^2

111 Falconer's heritability estimates (h^2) are sometimes not reported explicitly and I compute them from the reported concordances between MZ and DZ twins.

112 Twins Early Development Study (www.teds.ac.uk and http://www.iop.kcl.ac.uk/iopweb/departments/home/default.aspx?locator=336) is directed by Robert Plomin and involves all the twins born in England and Wales between 1994-1996 (~4000 pairs).

≈ 0.52 for Schonell-graded word reading test and $h^2 \approx 0.72$ for Schonell-graded word spelling test, $h^2 \approx 0.46$ for PIAT composite and $h^2 \approx 0.50$ for TOWRE. Fisher, Lai & Monaco (2003:59) report that concordances between MZ and DZ twins employing a strict definition of speech and language disorder are 70% versus 46% (giving $h^2 \approx 0.48$) and they rose to almost 100% versus 50% when the diagnostic criteria were broadened (giving $h^2 \approx 1.00$). Bishop (2003:S145), focusing on SLI[113], reports (from the literature) concordance rates (MZ vs DZ) of 0.86 vs 0.48 (giving $h^2 \approx 0.76$), 0.70 vs 0.46 (giving $h^2 \approx 0.48$) and 0.96 vs 0.69 (giving $h^2 \approx 0.54$), depending on the study. Bonneau, Verny & Uzé (2004:1214) report concordance rates of 85% (MZ) vs 50% (DZ) (giving $h^2 \approx 0.70$) and $h^2 \approx 0.73$, depending on the study.

Stromswold (2001:658) reports the results from 5 studies of written-language disorders, with concordance rates (MZ vs DZ) of 0.91 vs 0.45 (giving $h^2 \approx 0.92$) or 0.68 vs 0.43 (giving $h^2 \approx 0.50$) for dyslexia, 0.87 vs 0.33 (giving $h^2 \approx 1.08$) for dyslexia or problems with written language, 0.33 vs 0.29 (giving $h^2 \approx 0.08$) for Neale reading test, 0.35 vs 0.31 (giving $h^2 \approx 0.08$) for Schonell reading test, 0.50 vs 0.33 (giving $h^2 \approx 0.34$) for spelling and 1.00 vs 0.50 (giving $h^2 \approx 1.00$) for word blindness. The mean concordances for written-language disorders are thus 0.76 vs 0.40 (giving $h^2 \approx 0.72$) and the overall concordances (Stromswold, 2001) are 0.75 vs 0.42 (giving $h^2 \approx 0.66$). For spoken-language disorders, she reviewed 5 studies and the concordance rates (MZ vs DZ) are: 0.70 vs 0.46 (giving $h^2 \approx 0.48$) for SLI (strict criteria) and 0.94 vs 0.62 (giving $h^2 \approx 0.64$) for SLI (strict criteria), 0.81 vs 0.42 (giving $h^2 \approx 0.78$) for small vocabulary, 0.89 vs 0.55 (giving $h^2 \approx 0.68$) for SLI, 0.96 vs 0.69 (giving $h^2 \approx 0.54$) for poor composite language score, 0.98 vs 0.36 (giving $h^2 \approx 1.24$) for articulation problems, 0.83 vs 0.00 (giving $h^2 \approx 1.66$) for speech delay and 0.70 vs 0.50 (giving $h^2 \approx 0.40$) for language disorder in general. The average concordances (MZ vs DZ) are 0.84 vs 0.52 (giving $h^2 \approx 0.64$) and the overall polled concordances are 0.84 vs 0.52 (giving $h^2 \approx 0.72$).

The image created by these studies of language and speech disorders is that the heritability of their various aspects is generally very high ($h^2 > 0.50$)[114] and different for different aspects of language (spoken vs written language, articulation vs central processing, vocabulary vs

113Specific Language Impairment (SLI) is a complex language pathology; see below.

114For a comparison see Halliburton (2004:540, Table 13.4): min(h^2) = 0.00 (Red deer, number of offspring), max(h^2) = 0.92 (Humans, fingerprint ridge count), mean(h^2) = 0.43, median(h^2) = 0.40. For humans, h^2(height) = 0.65, h^2(schizophrenia) = 0.70, h^2(blood pressure) = 0.60 and h^2(IQ tests) = 0.50.

morpho-syntax). This supports the view that genetic factors play an important role in language pathologies and that different genetic factors are involved in different types of language pathology (Stromswold, 2001).

The heritability of normal aspects of speech and language is also high, but the number of studies dedicated to them is much more limited. For example, Plomin & Kovas (2005:595-597, Tables 1 & 2) report heritability estimates (h^2) of 0.22 for auditory repetition, 1.17 for non-word repetition and 0.48, 0.22 and 0.16 for various composite measures of language tests (in different studies). They also report heritability estimates for written-language measures of 0.18 for Neale accuracy reading, 0.44 for Neale comprehension reading, 0.19 for Schonell-graded word reading test, 0.53 for Schonell-graded word spelling test, 0.42 for a composite reading and spelling measure and 0.70 for TOWRE. Stromswold (2001:667, Table 5) reviews 8 studies concerning the heritability of vocabulary in normal populations and found h^2 estimates ranging from 0.02 for preferential looking at 14 and 20 months and expressive vocabulary at 14 months, to as high as 0.38 for expressive vocabulary at 24 months, 0.92 for the Stanford-Binet scale, 1.41 for Mehrabian vocabulary and 0.72 for WISC-R vocabulary. This implies a strong genetic component in normal vocabulary abilities, with "[d]ifferent factors [possibly] involved in the earliest stages of vocabulary acquisition compared to later stages [...]" (Stromswold, 2001:669). For phonology and articulation, "heritable factors account for 65% of the variance in children's phonemic abilities [...]" (Stromswold, 2001:671), with for example, the articulation of the phoneme /r/ being largely a result of genetic factors (Stromswold, 2001:673). In what concerns morpho-syntax Stromswold (2001:675, 676, Table 7), "genetic factors play a role in children's comprehension and production of syntax and morphology" (Stromswold, 2001:680). The estimated Falconer's heritabilities vary very much between different aspects (Stromswold, 2001:676, Table 7). These studies suggest that

> there are genetic factors for morphosyntax above and beyond the genetic factors that influence general nonverbal abilities [...] [and] the genetic factors that influence syntactic development are largely different from those that influence nonverbal cognitive ability, but [...] largely the same as those that affect vocabulary development (Stromswold, 2001:681).

The overall conclusion is that both in normal and pathological language and speech, genetic factors play an important role, but this role is a different function of the specific aspects concerned. The relationship between the genetic factors involved in normality and pathology

as well as the number of loci involved will be discussed in the following sections. It must be noted that Thompson *et al.* (2001) seem to support these findings through a study of the heritability of brain morphology and function, where they found that the structures of Broca's and Wernicke's areas are very heritable: for example, "[...] the asymmetry in language-related cortex was significant [...] in that Wernicke's and Broca's speech area displayed highly significant heritability on the left ($p < 0.0001$) but not on the right ($p > 0.05$)." (Thompson *et al.*, 2001:1255).

3.1.4. Beyond heritability part I: hunting genes, quantitative genetics and SLI

The heritability studies prove beyond any doubt that genetic factors play some role, sometimes even a major one, in both language pathology and normal development (Stromswold, 2001; Plomin *et al.*, 2001; Bishop, 2003; Fisher, Lai & Monaco, 2003; Felsenfeld, 2002; Plomin, Colledge & Dale, 2002). But how can we start looking for the actual genes involved in language?

The usual assumption when discussing the relationship between genotype and phenotype, in evolutionary contexts or otherwise, is that there is one (or a limited number of) gene(s) affecting a given trait (or limited number of traits) (Halliburton, 2004; West-Eberhard, 2003), an assumption carried into mathematical and computer modeling, either for technical/clarity reasons or because it is simply taken to be generally valid. But a thorough understanding of this relationship turns out to point to very complex interactions between both genes and environmental factors (Halliburton 2004; West-Eberhard, 2003; Plomin *et al.*, 2001), thus falsifying this simple assumption in most cases.

A *quantitative trait locus* (QTL) is a genetic locus affecting a quantitative trait; it corresponds to a small region of a chromosome and may encompass several tightly-linked genes (Halliburton, 2004:565). A *quantitative trait* (Halliburton, 2004:525) is a phenotype which displays a continuous range of variation and can be either a *true quantitative trait*, which have a continuous distribution (e.g., height, weight, etc.) or a *meristic trait*, which can have only (many) integer values (e.g., fingerprint ridge counts). A *threshold trait* is a closely related concept whereby an underlying quantitative trait gives rise to a binary classification

through the imposition of a cutoff value (e.g., most diagnostics of language pathologies). The reconciliation between Darwin's theory of natural selection and Mendelian genetics in the early twentieth century represented the birth of the New Synthesis and a crucial step was represented by the realization that quantitative traits can be accommodated by a Mendelian framework (Halliburton, 2004:7-15). They are usually controlled by many genes acting in concert and the environmental factors have significant effects on them (Halliburton, 2004:525).

To estimate the likely number of loci affecting a given trait, we can use the Castle-Wright estimator, which evaluates the phenotypic difference between two inbred lines crossed together: F_1 is the first generation of crosses between the inbred lines and F_2 is the second generation produced from F_1 individuals. The phenotypic variation in F_2 is greater than in F_1 because now different alleles can segregate. An underestimate of the number of loci affecting this trait is given by the *effective number of loci*:

$$n_e = (M_1 - M_2) / 8(V_{F2} - V_{F1})$$

where M_1 and M_2 represent the average value of the trait in the two inbred lines and V_{F1} and V_{F2} represent the trait variance in F_1 and F_2 respectively (Halliburton, 2004:565). The number of QTLs estimated using this technique varies from 5 (skin color in humans) to as many as 157-485 (pupa weight in the red flour beetle) and 164 (litter size in mice) (Halliburton, 2004:566, Table 13.11), but it is usually of 20 or fewer (Halliburton, 2004:565).

To actually map the QTL to positions on the chromosomes, one has to use the fact that the *genetic linkage*[115] between two loci is not constant, but varies with their physical position on the chromosomes. The unit of measurement of genetic distance is the *centiMorgan* (cM): 1cM represents the distance between two loci for which there is a 0.01 probability of recombination in one generation (Jobling, Hurles & Tyler-Smith, 2004:36). Two loci separated by 50cM or more are essentially unlinked, being randomly assorted. On average, at the scale of the entire human genome, 1cM represents approximately 1Mbp (Jobling, Hurles & Tyler-Smith, 2004:37). In *association studies* (Plomin *et al.*, 2001:369; Jobling, Hurles & Tyler-Smith, 2004:447; Halliburton, 2004:565; Bishop, 2002:316-318), a population of

115The genetic linkage reflects the fact that some loci tend to be transmitted together (Jobling, Hurles & Tyler-Smith, 2004:36).

individuals displaying a given trait (or high values of it) and another population of individuals without the trait (or low values of it) are compared, so that specific genetic markers associated with the trait (presence or values) are identified (Plomin *et al.*, 2001:369; Jobling, Hurles & Tyler-Smith, 2004:447). There are many such markers mapped[116] and an association study will try to find statistical correlations between such markers and the considered trait. After the most promising markers have been identified, the positions of the actual loci are inferred from the positions of the markers. A good overview of this complex process, including the main two hypotheses: the common disease/common variant[117] and the genetic heterogeneity model[118], is Jobling, Hurles & Tyler-Smith (2004:452-459).

A very interesting case from a linguistic point of view is represented by SLI. Specific Language Impairment (OMIM 602081) is defined partially by exclusion (Bishop, 2003:S143) as

> a disorder in the development of language despite adequate educational opportunity and normal intelligence. A diagnosis requires a significant discrepancy between the child's verbal and nonverbal abilities in the absence of any additional disorders that might underlie the language problems (e.g., hearing loss, mental retardation, and autism) (SLIC, 2004:1225).

The phenotype is complex and probably comprises more subtypes (van der Lely, 2005; The SLI Consortium, 2004:1225). Moreover, there is still argument concerning its status (Bishop, 2003:S153): some regard it as qualitatively different from normality (i.e. a distinct phenotype) (e.g., O'Brien *et al.*, 2003) while some regard it as the tail end of the normal distribution of language abilities (e.g., Plomin, Colledge & Dale, 2002:420). Through association studies, SLI seems reliably connected to the 7q31 region (O'Brien *et al.*, 2003), a region onto which autism also maps (O'Brien *et al.*, 2003:1536), but distinct from *FOXP2* (also in the same region). SLI is also mapped to markers on regions 16q24 and 19q13 (Fisher, Lai & Monaco, 2003:72; van der Lely, 2005:53; Plomin, Colledge & Dale, 2002:423), but the exact genes involved are unknown for the moment.

Also, there is controversy about the "core deficit" of SLI, with some arguing for a language-

---

116See for example, the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606) (September 2006).

117Common disease is influenced by a small number of susceptibility alleles at each locus (Jobling, Hurles & Tyler-Smith, 2004:452).

118Common disease is influenced by many rare alleles at many loci (Jobling, Hurles & Tyler-Smith, 2004:454).

specific disorder (e.g., van der Lely's *G-SLI*, 2005), while others support a view of SLI as a more general, not-so-language-specific disorder (Bishop, 2002; Newbury, Bishop & Monaco, 2005). The last type of theories argues for a complex interaction between two different deficits:

> [w]e were left [after analyzing the genetic data] with the intriguing puzzle that SLI was associated with two impairments – one in phonological short-term memory, and the other in auditory processing – but these were not simply different indices of the same thing (Bishop, 2002:323)

and concludes that "[...] the simplest conclusion is that underlying impairments in auditory processing and phonological short-term memory act as additive risk factors for language impairment" (p. 324). This can be represented graphically by a *double hit model* (Figure 16), where the environment influences much stronger on the auditory deficit and genes on the phonological short-term memory deficit (Bishop, 2002).

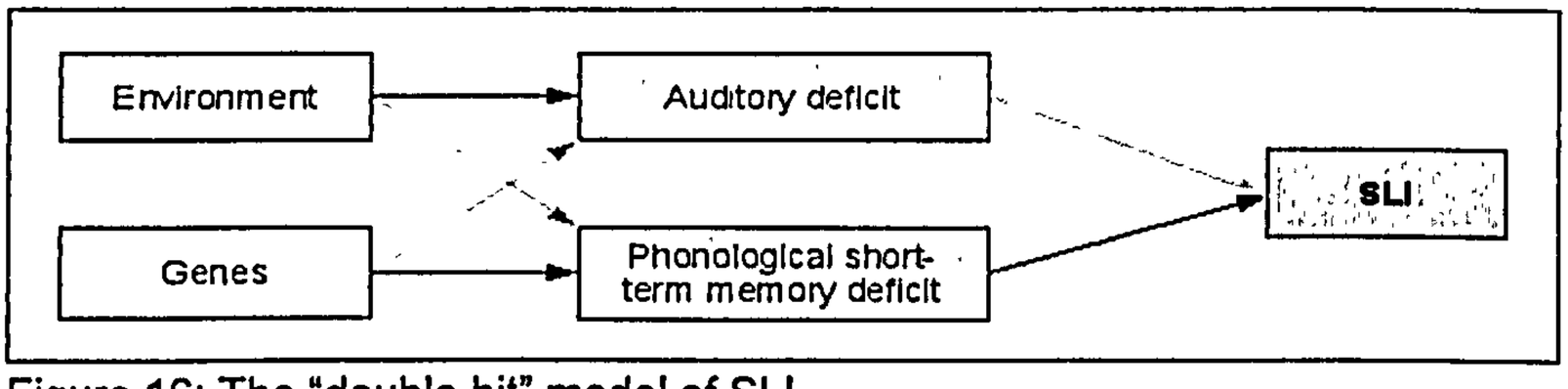

Figure 16: The "double hit" model of SLI.

There are two underlying impairments which have to be present simultaneously to produce the SLI phenotype: an auditory deficit (mostly determined by the environment) and a phonological short-term memory deficit (mostly under genetic influences). The strongest influences are represented by black arrows while the weakest ones by gray arrows. The most important factor in the etiology of SLI is the strongly genetically influenced deficit on phonological short-term memory. Adapted from Bishop (2002:324), Figure 7.

This model is further supported and refined by Newbury, Bishop & Monaco (2005): the NWR (non-word repetition tasks, a test of phonological short-term memory capacity) deficit is highly heritable, while deficits in non-verbal auditory tasks (like pure tone discrimination tests) appear to be heavily influenced by environment. This enriched risk-factor model for SLI can be summarized as: "[r]ather than being different manifestations of the same underlying disorder, auditory, phonological and morphosyntactic deficits have distinct causes, and each deficit increases the probability that clinically significant SLI will result" (Newbury, Bishop & Monaco, 2005:530), seemingly also supporting van der Lely's (2005) G-SLI hypothesis. This is represented in Figure 17 below. Even more interesting is the fact that the authors propose that the chromosome 19 locus (19q13, see above) is likely to

generally influence language-related processes, but the chromosome 16 locus (16q24, see above) is more specific to phonological short-term memory (Newbury, Bishop & Monaco, 2005:531).

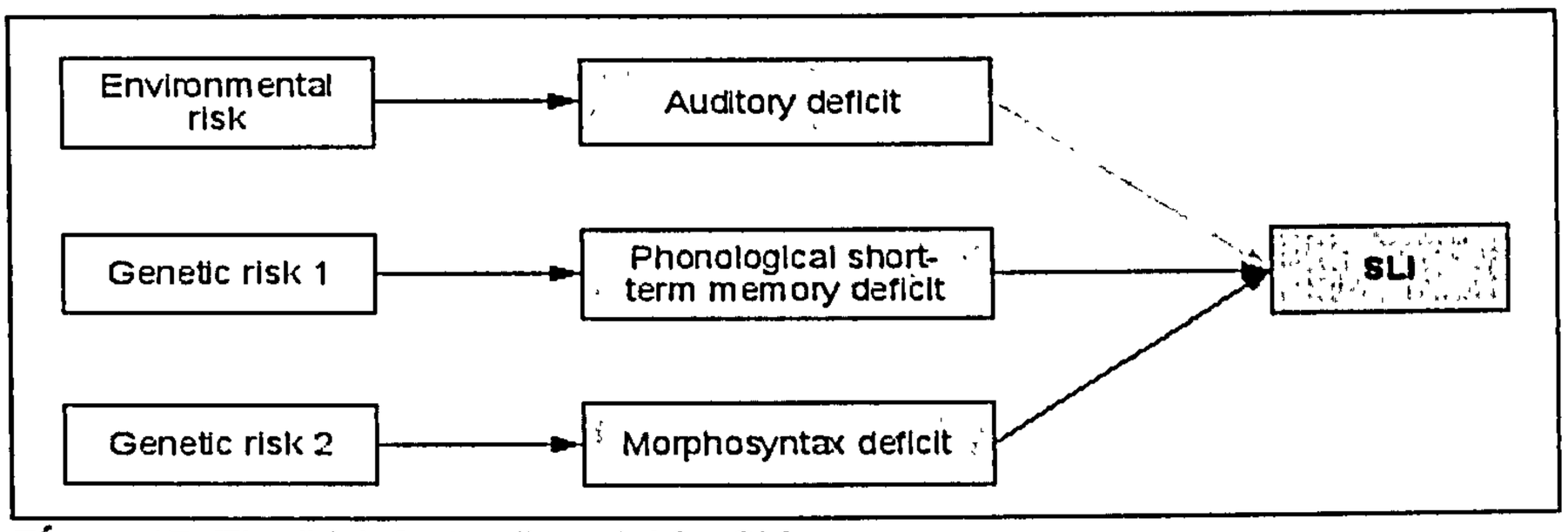

Figure 17: The "risk factors" model for SLI.
See main text for details. Adapted from Newbury, Bishop & Monaco (2005:530), Figure 3.

To conclude, SLI seems a case where QTL methodology is appropriate, as this deficit is very complex, involving many loci on several chromosomes, all involved in producing the clinical deficit. That SLI is indeed not a simple matter is further confirmed by a specific subtype, strongly associated with the *FOXP2* gene and involving a different methodology for finding genes.

3.1.5. Beyond heritability part II: hunting the *FOXP2* gene

The literature concerning *FOXP2* is already large and growing fast, as more research is focused on this "star" gene and more speculation is entertained each time new data emerge. Already, almost any discussion concerning human and/or language evolution makes at least a passing reference to this gene and its putative implications (e.g., Bickerton, *in press*; Pinker & Jackendoff, 2005; Corballis, 2004).

The story started in the '90s, as the British KE family came to the attention of both the scientific community and the public (Gopnik & Crago, 1991). The pedigree of this family, probably the currently best-known genealogy among linguists, is reproduced in Figure 18, and depicts a three-generations family with half the members (15 out of 31) affected by a complex pathology, involving speech and language (Hurst *et al.*, 1990). The actual

phenotype is still debated (Bishop, 2002; Fisher, Lai & Monaco, 2003; Lai *et al.*, 2001; Vargha-Khadem *et al.*, 1998; Lai *et al.*, 2003; Marcus & Fisher, 2003; Liégeois *et al.*, 2003; Watkins, Dronkers & Vargha-Khadem, 2002; Watkins *et al.*, 2002) but it seems that the picture is very complex and concerns (Fisher, Lai & Monaco, 2003:64-66):

- *articulatory problems*: affected individuals have troubles with coordinating complex oro-facial movements, not resulting from impairment in simple oro-facial movements nor from abnormalities in facial muscles;
- *cognitive impairments*: the average non-verbal IQ of the affected members is lower than that of the non-affected members but there is a large overlap between them (Marcus & Fisher, 2003:258), while the verbal IQ is significantly affected in all of its separate components;
- *language impairments*: a set of problems with both spoken (expressive and receptive) and written language are detected, and the disorder affects both the comprehension and production components of grammar (understanding complex sentences, inflectional and derivational morphology).

The disorder is classified as *developmental verbal dyspraxia* (OMIM 602081) and included in the SLI category.

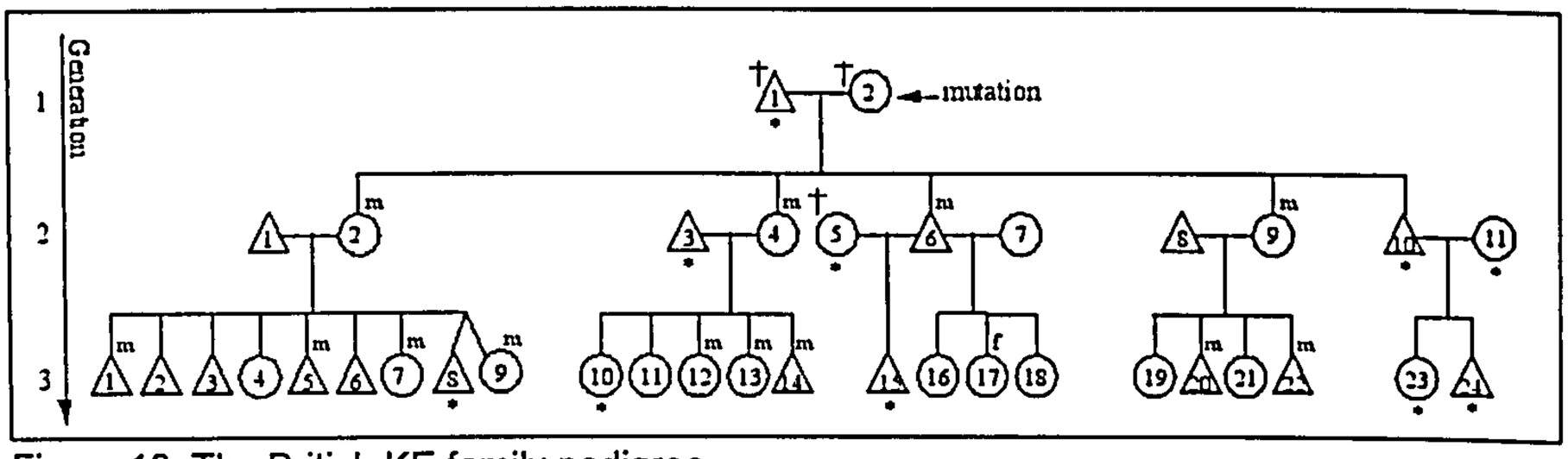

Figure 18: The British KE family pedigree.

Adapted from Fisher, Lai & Monaco, 2003:63, Figure 1; Bishop, 2002:312, Fig. 1; Lai et al., 2001:519, Figure 1. The triangles represent males and the circles females. The generations flow from top (1) to bottom (3). Individuals are numbered starting with 1 in each generation. Affected individuals are light gray and "m" means that they inherited the pathology from the mother and "f" from the father. Dead individuals are marked with a cross. Individuals unavailable for genetic assessment have an star (*). Individuals 8 and 9 in generation 3 are non-identical twins.

What made the KE family and its peculiar pathology an instant focus of research is the fact that, as can be seen from their pedigree (Figure 18), there is a clear genetic pattern of

transmission of the disease, following a *dominant autosomal* model of inheritance (Hurst *et al.*, 1990; Bishop, 2002; Lai *et al.*, 2001). The grandmother (individual 2, generation 1) is the one in which the mutation appeared (there is no known history of language-related problems beyond her, Fisher, Lai & Monaco, 2003:73). The mutation is fully penetrant and strongly correlated with the pathology, which makes it a textbook case of pedigree study (Lai *et al.*, 2001; Fisher, Lai & Monaco, 2003:63). An earlier study (Fisher *et al.*, 1998), localized the locus, named *SPCH1* (OMIM 602081), on the 7th chromosome (7q31), using the co-segregation of the pathology with a number of genetic markers. This region was further refined by Lai *et al.* (2000) to approximately 6.1Mb. A spectacular progress was made when an unrelated individual (CS), suffering from a similar pathology, turned out to have a *de novo* translocation[119] affecting this region (7q31), which allowed the precise identification of the gene (Lai *et al.*, 2001).

It turned out that this gene is a member of the *Forkhead box* (Fox) family of genes (Lai *et al.*, 2001). There are at least 40 members of this family in humans, acting as *transcription regulators*[120], through highly conserved specialized regions (80-100AA long), the *DNA-binding domains*, that interact with the target genes' promoters[121] (Scharff & White, 2004:329). The family's name derives from the fork-like head phenotype of the Drosophila embryos produced by the first mutant of a FOX protein discovered (Scharff & White, 2004:329). This family is subdivided into subfamilies, distinguished by letters (*FOXA* to *FOXQ* so far); the *FOXP* subfamily has four members (*FOXP1* to *FOXP4*) (Scharff & White, 2004:329). *FOXP1* and *FOXP2* are very similar and it seems that they interact during expression, forming hetero-dimers (Scharff & White, 2004:330; Vargha-Khadem *et al.*, 2005:135; Teramitsu *et al.*, 2004:3153). The nomenclature across species is complex (Teramitsu *et al.* 2004:3152): the genes are italic (*FOXP2*) while proteins are not (FOXP2); human forms use uppercase (FOXP2), murine (mouse) forms are lowercase (Foxp2) while for other species they are a combination of lower- and uppercase (FoxP2).

The *FOXP2* gene is composed of 23 exons (Lai *et al.*, 2001; Bruce & Margolis, 2002) and

---

119The translocation is t(5;7)(q22;q31.2), involving thus chromosomes 5 and 7, and was not present in the proband's parents.

120They can alter the production of mRNA of certain genes (increasing or decreasing it) (Scharff & White, 2004:329).

121The region of a gene allowing it to be transcribed into mRNA by a RNA polymerase (Lewin, 2004).

can be alternatively spliced. The mutation associated with the KE pathology is a *missense G-to-A transition*[122] in exon 14, producing an *arginine-to-histidine (R553H) substitution in the forkhead DNA-binding domain* (Lai *et al.*, 2001:520-521). This mutation does not represent a polymorphism in human populations and is inferred to disrupt the DNA-binding properties of FOXP2 (*Lai et al.*, 2001). Also, the translocation in CS occurred in the intron between exons 3b and 4 and heavily disrupted the structure of FOXP2 (Lai *et al.*, 2001:520). Recently, MacDermot *et al.* (2005) have reported a new mutation in exon 7, a *C-to-T transition* giving a *stop codon*[123] *at position 328* (R328X) (MacDermot *et al.*, 2005:1076) and associated with a speech and language pathology similar to the KE and CS cases in a proband, his affected sibling and their mother (MacDermot *et al.*, 2005:1076). This mutation is also not a polymorphism in human populations (MacDermot *et al.*, 2005:1076).

Evolutionary and comparative studies of *FoxP2* (Webb & Zhang, 2005; Enard *et al.*, 2002; Zhang, Webb & Podlaha, 2002; Scharff & Haesler, 2005; Teramitsu *et al.*, 2004; Haesler *et al.*, 2004; Shu *et al.*, 2005) have found that it is very conserved across taxa: it belongs to the 5% most conserved genes in a human-rodent comparison involving 1880 genes (Enard *et al.*, 2002:869). This comparison of primate and mouse FoxP2 (Enard *et al.*, 2002), allowed the identification of the fact that the human and mouse proteins differ at only three amino-acid positions, and two of them are specific to humans, both in exon 7 (threonine-to-asparagine at 303 and aspargine-to-serine at 325) (Figure 19). Zhang, Webb & Podlaha (2002:1829) found that one of the two human-specific substitutions (aspargine-to-serine at 325) also occurs in the order *Carnivora*, independently from *Homo sapiens*, "suggesting that this substitution alone is not sufficient for the origin of speech and language" (Zhang, Webb & Podlaha, 2002:1829), which is to be expected given that the single catastrophic mutation theories of language origins are very unlikely (see below).

The two human-specific substitutions seem fixed in the human population (Enard *et al.*, 2002:870; Zhang, Webb & Podlaha, 2002:1829-1830) and an evolutionary analysis of the substitution rates and polymorphism levels (Enard *et al.*, 2002; Zhang, Webb & Podlaha, 2002) suggests that the human allele was exposed to strong selection, either *background*

122Missense or nonsynonymous mutations change the AA sequence of the resulting protein. A transition replaces a purine with another purine (A ↔ G) or a pyrimidine with another pyrimidine (C ↔ T) (Jobling, Hurled & Tyler-Smith, 2004:47).

123A stop codon provokes the termination of the translation process (Jobling, Hurles & Tyler-Smith, 2004:26).

selection[124] or *selective sweeps* (Zhang, Webb & Podlaha, 2002:1830), while a relaxation of selective constraints cannot be ruled out, but is most unlikely given the deleterious effects of the known mutations (Enard *et al.*, 2002:870; Zhang, Webb & Podlaha, 2002:1830).

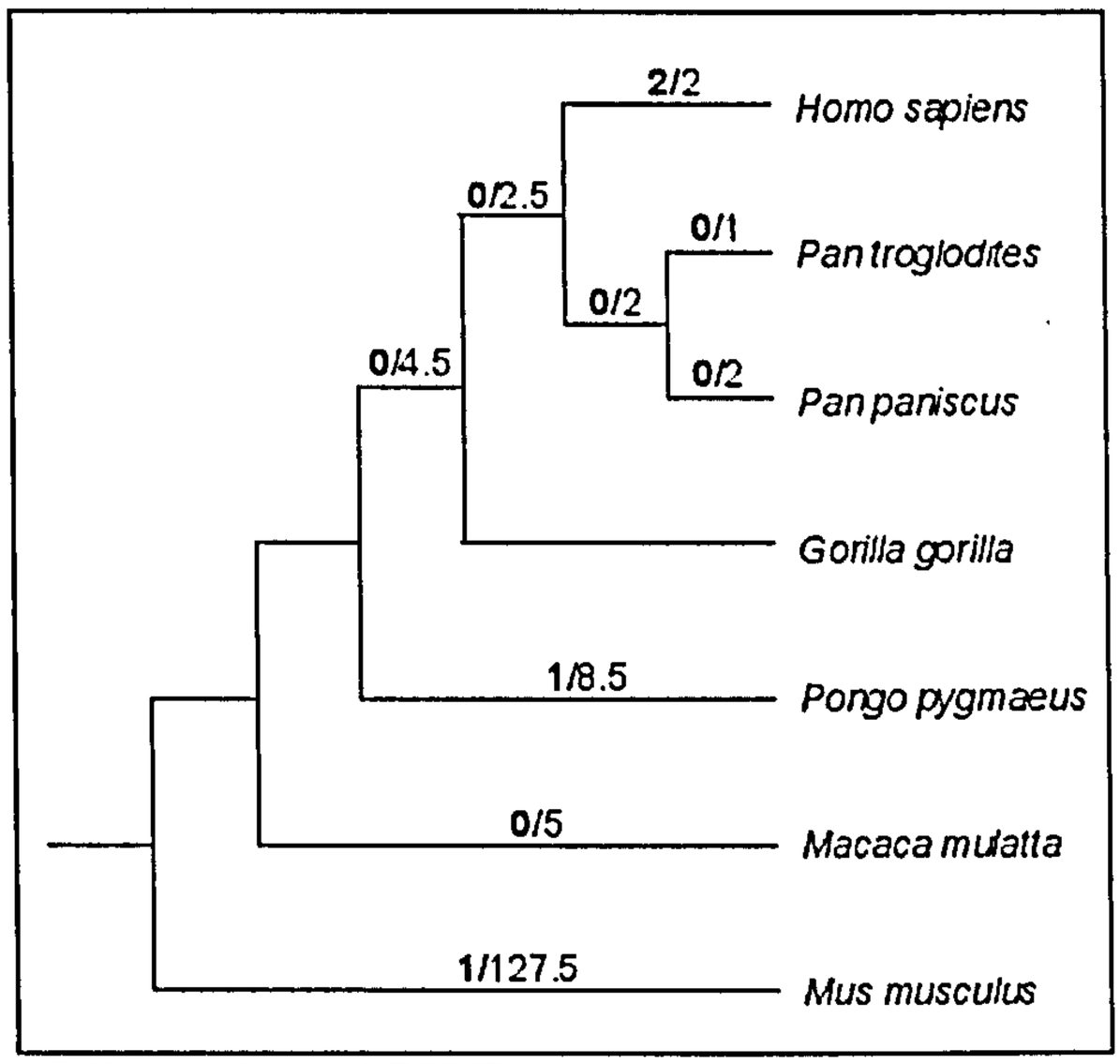

Figure 19: Evolutionary tree of *FOXP2*.

*It shows both non-synonymous (**bold**, first number) and synonymous (regular font, second number) substitutions in primate and rodent lineages. Adapted from Zhang, Webb & Podlaha (2002:1829, Figure 4) and Enard et al. (2002:871, Figure 2).*

Estimating the age of fixation of this allele is fraught with difficulties. Enard *et al.* (2002:871) used a constant-size panmictic model to estimate this date and obtained that the most likely date is 0ya, with a 95% CI of (0, 120kya) but argued that if a population growth soon follows this fixation, the date may be pushed back

> by at most the time since the onset of human population growth, some 10,000-100,000 years ago. In any case, our method suggests that the fixation occurred during the last 200,000 years of human history, that is, concomitant with or subsequent to the emergence of anatomically modern humans. This is compatible with a model in which the expansion of modern humans was driven by the appearance of a more-proficient spoken language (Enard *et al.*, 2002:871).

Zhang, Webb & Podlaha, (2002:1831) are much more vague about the estimate: "[...] the sweep would have occurred no earlier than 5000 generations, or ~100,000 years, ago. This estimate is within the wide window of 40,000 years to 4 MYA during which human languages are believed to have emerged" (Zhang, Webb & Podlaha, 2002:1831).

124Defined as purifying selection on deleterious mutations in tightly linked exons (Zhang, Webb & Podlaha, 2002:1830).

There are many problems with these estimates. First, the Enard *et al.* (2002) date uses a highly unlikely demographic model (constant size panmictic population), modified by a discussion of an exponential growth effect. The age of this effect is taken to be 10-100kya using Wall & Prezworski (2000) as a reference, while, in fact, this paper does not provide any firm estimate but a discussion of various demographic scenarios more or less compatible with nuclear marker signals. The most straightforward interpretation of their findings is simply that *we don't know yet when exactly the selective sweep occurred*: it might be ancient (>100kya) or very recent. Cases such as discussed in Mekel-Bobrov *et al.* (2005), Evans *et al.* (2005) and Voight *et al.* (2006) prove that ongoing selective sweeps happen in humans and it is quite possible in principle that the fixation of the current human *FOXP2* allele is a fairly recent phenomenon. Second, even if a more reliable estimate of the date of fixation of this allele would be available, it would certainly be hard to interpret it in terms of language evolution. Given that there are two human-specific replacement substitutions and there is currently no data on the effect of reversal mutations at these positions on language, it is entirely undecidable for the moment which of the many possible scenarios is more probable. It is possible, for example, that each mutation in turn provoked a selective sweep but only the first one is connected to language while the second was involved in something else, or, that both mutations are involved in language and there were two independent selective sweeps, or, that only the last mutation in combination with the neutral preexisting first one became positively selected, just to list a few possibilities[125]. Thus, any simplistic claims that *FOXP2 proves* language to be recent or connected to modern human origins and expansion are simply unsupported.

To complicate matters even further, studies of *FoxP2* in avian species with and without learned song and non-human vocal-learning mammals (Webb & Zhang, 2005; Teramitsu *et al.*, 2004; Scharff & Haesler, 2005; Haesler *et al.*, 2004; Shu *et al.*, 2005), have tended to find that the exon 7 is conserved in both song-learning and non-song-learning birds, while the whales, bats and humans[126] do not share any amino-acid changes in exon 7. Interestingly, "whales and dolphins share three amino-acid substitutions while their closest relative, the hippopotamus, is identical to mouse. Notably, the human-unique substitution (T303N) was flanked by two changes in both whale and dolphin (S302P and T304A)" (Webb & Zhang,

125"First" and "second" do not refer to positions on the chromosome but to their (unknown) temporal precedence.

126Whales, dolphins, bats and humans are considered to be the only vocal-learning mammals by Webb & Zhang, (2005:214), but not everybody agrees.

2005:214 & Figure 2). This would seem to imply that somehow changes at position 303 and immediate neighbors are connected to vocal-learning in mammals (this can be accommodated by the bat-specific change S298L but the trouble seems to be the tapir-specific change T304A, identical with the whale and dolphin at that position). Clearly, a better understanding of the protein's function is required before meaningful evolutionary generalizations can be made. As Haesler *et al.* (2004:3173) put it:

> [t]aken together, we conclude that the striking conservation of the *FoxP2* gene sequence and overall brain expression pattern in avian, reptilian, and mammalian brains, regardless of whether they learn to vocalize or not, confirms that *FoxP2* has a more general role than to enable vocal learning. *FoxP2* could be an ancient transcription factor involved in shaping cerebral architecture, perhaps via restriction of certain neuronal lineages (Haesler *et al.*, 2004:3173)

Concerning specifically song-learning birds, *FoxP2* expression seems to correlate with song-related brain areas (spatially) and with song plasticity (temporally) (Haesler *et al.*, 2004; Scharff & Haesler, 2005). *Foxp2* silencing in mice (Shu *et al.*, 2005) produces a very interesting pattern of disruption in the heterozygote, including the ultrasonic vocalizations with function in stress response: "[...] the frequency of occurrence of ultrasonic vocalizations is selectively impaired in the knockout and heterozygous mice [...] [while] the apparatus necessary for the production of vocalizations, including the neural control, in the vocal tract, and brainstem, is normal" (Shu *et al.*, 2005:9647), but one should be very careful when generalizing from mice to man (Shu *et al.*, 2005). It is interesting to note, however, that besides the parallel impairment in vocal signaling, the same pattern of homozygous lethality versus heterozygous subnormal functioning is maintained, supporting a *quantitative deficit hypothesis*, whereby having half the normal quantity of functioning FOXP2 produces non-lethal developmental alterations (Lai *et al.*, 2001). As the review of Scharff & Haesler (2005) conclude, after summarizing the parallel brain-expression pattern of *FoxP1* and *FoxP2* across species and silencing deficits in mouse,

> [t]he original suspicion that FoxP2 would be primarily involved in control of oro-facial muscles and, thus, would be only peripherally interesting for understanding neural substrates for speech and language, is not supported by the gene expression and mouse KO data. Instead, the strong expression of *Foxp2* in cerebellar and basal ganglia circuits points towards functions that include sensory-motor integration important for sequenced behaviors and procedural learning (Scharff & Haesler, 2005:699-700).

Returning to humans, after the initial announcement (Lai *et al.*, 2001) of *FOXP2* disruption

being involved in SLI, a plethora of studies have tackled related areas of deficit. Thus, it is acknowledged that *autism* (OMIM 209850) has associated language impairments very much like those found in SLI and that one of the loci involved maps to 7q31-32 (Li *et al.*, 2005; Newbury *et al.*, 2002): thus, a natural question concerns the possible involvement of *FOXP2* disruption in autism. Unfortunately, both studies conclude that "the SPCH1 [*FOXP2*] and AUTS1 [the locus on 7q31 involved in autism] are attributable to different genes that, coincidentally, lie in similar positions on chromosome 7q" (Newbury *et al.*, 2002:1324) and that *FOXP2* is probably not involved in autism. *Schizophrenia* (OMIM 181500) was also considered as possibly involving disruptions in *FOXP2*, especially because, besides language-related impairments, some language-evolution theories link the linguistic capacity to susceptibility to schizophrenia (e.g. Crow, 2002c). Unfortunately again, Sanjuan *et al.* (2005), fail to find any connection between *FOXP2* and schizophrenia. But is *FOXP2* a susceptibility gene for more common forms of SLI? It is clear that the KE, CS (Lai *et al.*, 2001) and the family reported in MacDermot *et al.* (2005) have a very special form of SLI for which disruption of *FOXP2* is the causal explanation, but Newbury *et al.* (2002) conclude that "[...] it would appear that the role of FOXP2 in speech and language disorders does not generalize to more common and genetically complex forms of language impairment" (Newbury *et al.*, 2002:1324), conclusion supported also by many others (Bishop, 2003; Scerif & Karmiloff-Smith, 2005; Marcus & Fisher, 2003; Felsenfeld, 2002; O'Brien *et al.*, 2003; Plomin, Colledge & Dale, 2002; Fisher, Lai & Monaco, 2003).

The phenotypic effects of disruptions in *FOXP2* have also been analyzed at the neural level, both in adult and developing brains. Vargha-Khadem *et al.* (1998) were the first to undertake a brain functional and structural study of the affected members of the KE family, finding sites of bilateral pathology in the basal ganglia (including reduced volume of the caudate nucleus), affecting cortical motor areas relevant for speech and language (Vargha-Khadem *et al.*, 1998:12695, 12700). Watkins *et al.* (2002) performed MRI analyses of the affected and unaffected members of the KE family and a matched control group, and found significant differences, especially in the caudate nucleus. Later, Liégeois *et al.* (2003), conducted an fMRI study involving the same affected members of the KE family, and found that they seem affected not in fluency on semantic retrieval as such, but in rapidly selecting items from the semantic memory (Liégeois *et al.*, 2003:1233) and that they show high atypical activation patterns during linguistic tasks, with underactivation in Broca's area and other

cortical and sub-cortical regions (Liégeois *et al.*, 2003:1234). They conclude:

> [t]he *FOXP2* gene may therefore have an important role in the development of a putative frontostriatal network involved in the learning and/or planning and execution of speech motor sequences, similar to that involved in other motor skills (Liégeois *et al.*, 2003:1234).

A developmental study was reported in Lai *et al.* (2003), where temporal and spatial patterns of *FOXP2* expression in mouse and human developing brains is analyzed. They found that these patterns are very conserved in mouse and human and that there is no human-specific expression site in the developing brain (Lai *et al.*, 2003:2461); that there is a high concordance between the sites expressing *FOXP2* during development and those affected in *FOXP2*-deficient individuals (Lai *et al.*, 2003:2460). These sites involve mainly neural structures implicated in motor control (basal ganglia, thalamus, inferior olives, cerebellum) (Lai *et al.*, 2003:2460). *FOXP2* expression in the developing brain is neither uniform/diffuse nor strictly circumscribed, but it shows restricted expression in related brain areas; moreover, as development progresses, its expression is refined within those areas (Lai *et al.*, 2003:2458). A recent review (Vargha-Khadem *et al.*, 2005) proposes a neural circuit involved in speech and language showing processes affected by *FOXP2* (Vargha-Khadem *et al.*, 2005:136, Figure 4).

So, what does this gene tell as about language and its evolution? First of all, the deficit it produces is very special and not representative of more common types of language deficits, which are probably under combined polygenic and environmental control. Second, the date of its selective sweep and fixation in humans does not tell us anything about the evolution of language or its form before the allele's fixation. Third, it seems that exon 7 is somehow related to vocal learning in mammals but not in birds (other exons might behave differently, while *FoxP1* might also be relevant, Teramitsu *et al.*, 2004) but there is no interpretation possible for the moment. Fourth, it seems very probable that the deficit is related to motor control learning and not oro-facial movements. *FOXP2* represents, thus, a very interesting research avenue with a huge potential for language and speech, but probably not a major player in the language evolution arena.

> The striking conservation of the *FoxP2* gene sequence and overall brain expression pattern in reptilian and mammalian brains and in the brains of both song-learning and non-song-learning birds indicates that *FoxP2* has a more general role than to specifically enable vocal learning. *FoxP2* could be an ancient

> transcription factor primarily involved in setting up and maintaining subtelencephalic and striatal sensory and sensory-motor circuits, *creating a permissive environment upon which vocal learning can evolve* if other circumstances/factors come into play (Scharff & White, 2004:342, *italics* mine).

3.1.6. Beyond heritability part III: genes, abilities and disabilities

The fact that genetic factors account for an important proportion of variance in normal and pathological populations does not say anything about the relationship between these populations, nor about the structure of the pathologies. The remaining open questions concern the *genetic links between disabilities and abilities* (are they qualitatively or quantitatively different?), the *genetic links within disabilities/abilities* (are they homogeneous or composed of different sub-pathologies?) and the *genetic links between disabilities/abilities* (is there co-morbidity?) (Plomin & Kovas, 2005:592; Plomin *et al.*, 2001:150, Box 8.1; Stromswold, 2001:655).

It might be possible that disabilities represent just *the low end of the normal range of variation*, *quantitatively* different from abilities and influenced by the same genetic factors; the alternative is that they are *distinct entities*, *qualitatively* different from abilities and influenced by different genetic factors (Plomin & Kovas, 2005:592; Stromswold, 2002:655; Plomin *et al.*, 2001:150, Box 8.1). *DeFries-Fulker (DF) extremes analysis* (DeFries & Fulker, 1998; Plomin & Kovas, 2005:592; Stromswold, 2002:655; Plomin *et al.*, 2001:150, Box 8.1) was developed in order to compare dichotomous diagnoses of disability (concordance data) with continuous scores of ability (correlation data), using multiple regression (Tabachnick & Fidell, 2001:111-176) (Figure 20).

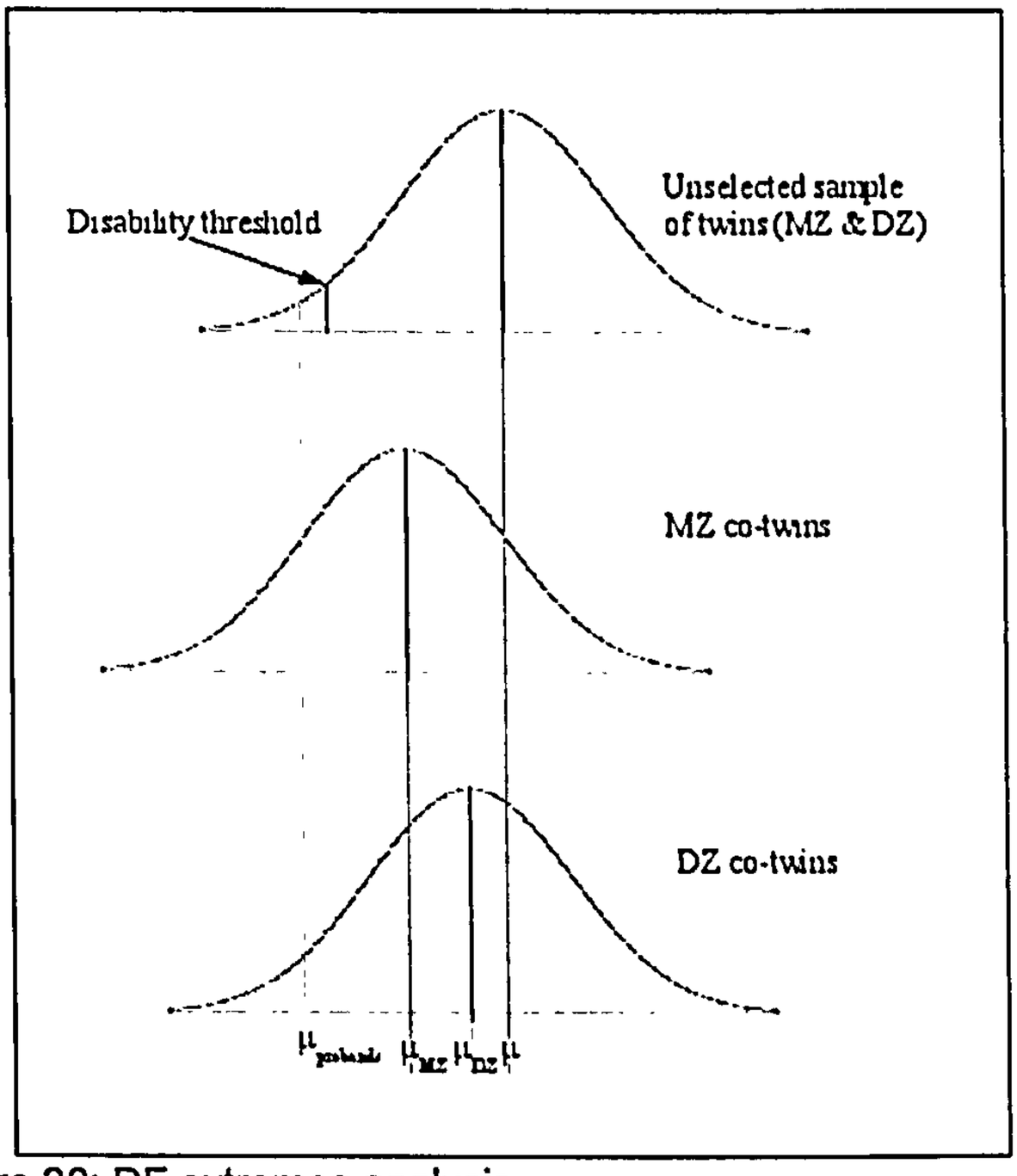

Figure 20: DF extremes analysis.

Adapted from Plomin et al., 2001:150, Box 8.1 and Plomin & Kovas, 2005:593, Figure 1. See text for details.

A population of MZ and DZ twins has an average μ for the trait of interest, while the affected population (the probands) has an average $\mu_{probands}$ for the same trait. If the trait is heritable, it is expected that the MZ twins of the probands will have more similar scores to the probands themselves than the DZ twins (DZ twins of probands will regress more towards the mean of the population). If μ_{MZ} and μ_{DZ} are the averages of MZ and DZ co-twins, then it is expected that:

$$\mu_{probands} \leq \mu_{MZ} \leq \mu_{DZ} \leq \mu$$

which would suggest that genetic factors contribute to the mean difference between the disabled probands and the population (Plomin *et al.*, 2001:150, Box 8.1). Standardized scores (de Vaus, 2002:109) can be fitted to a regression equation:

$$c = b_1 p + b_2 r + a$$

where c is the predicted score for the co-twin, p is the proband's score, r is the coefficient of genetic similarity between twins[127], a is the regression constant, b_1 is the partial regression of co-twin's score on the proband's score[128] and b_2 is the partial regression of co-twin's score on

1271.0 for MZ and 0.5 for DZ twins (Table 3, Section 3.1.2).
128Representing how similar the two twins are independent of zygosity (which is addressed by b_2).

the genetic relatedness[129] (Stromswold, 2001:655, Plomin & Kovas, 2005:593). If b_2 is significant, then there is a significant heritable effect: it represents the genetic contribution to the phenotypic mean difference between the probands and the population (Plomin & Kovas, 2005:593). The *group heritability* is estimated as:

$$h^2_g = b_2 / (\mu - \mu_{probands})$$

and represents a measure of the heritability of the trait in a population affected by a disorder (Stromswold, 2001:656), in a way quantifying the proportion of the difference between the affected and unselected means which can be accounted for by genetic factors (Plomin *et al.*, 2001:151, Box 8.1). It must be distinguished from h^2, which measures the heritability of the trait in the entire population. If there are genetic factors influencing only the variance of the affected population (the probands) but not the variance in the general ("normal", unaffected) population, then $h^2_g > h^2$ (Stromswold, 2001:656); an extreme example of genetic processes totally different for pathology and normality is the case of a single-gene disorder that contributes little to the normal variation, and which will have $h^2_g = 0$ (Plomin & Kovas, 2005:594).

Group heritabilities for language and speech (Plomin & Kovas, 2005:595, Table 1; 596; Stromswold, 2001) tend to vary depending on the component tested, but are generally substantial. For example, $h^2_g \approx 1.25$ for phonological short-term memory (Stromswold, 2001:660), $h^2_g \approx 0.11$ for pure tone repetition (Stromswold, 2001:661) while $h^2_g \approx 1.17$ for nonsense word repetition (Stromswold, 2001:661). Plomin & Kovas (2005:595, Table 1) also report $h^2_g \approx 0.45$ and $h^2_g \approx 0.37$ for various composite language tests. The overall conclusion of these studies seems to be that there are generally moderate estimates of h^2_g for language abilities and disabilities (an average of 0.47 for language ability) and that "[...] DF extremes group heritability is similar to liability heritability of disability and individual differences heritability of ability, suggesting *strong genetic links between language disability and ability*" (Plomin & Kovas, 2005:599, *italics* mine) This suggests that many language and speech pathologies are actually just the low end of the normal range of variation, quantitatively, and not qualitatively, different from this "normality". This, again, highlights the atypicality of *FOXP2* heterozygous deficits.

The second question concerns the homogeneity of language and speech disabilities/abilities,

129And equal to $2*(\mu_{DZ} - \mu_{MZ})$.

i.e., to what extent are the same genetic factors influencing the different aspects of the same disability? (Plomin & Kovas, 2005:600). *Multivariate genetic analysis* (Gillespie & Martin, 2005) attempts to "decompose the co-variance between traits into genetic and environmental sources of covariance" (Plomin & Kovas, 2005:600). One of the fundamental concepts is the *genetic correlation*, defined as "the extent to which genetic effects on trait *X* correlate with genetic effects on trait *Y* regardless of the heritabilities of *X* and *Y*" (Plomin & Kovas, 2005:600) and which can be interpreted as the probability that a gene influencing trait *X* will also influence trait *Y*[130]. Let's denote the genetic correlation between X and Y as g_{XY}; if g_{XY} = 1.0, then the same genes affect both *X* and *Y*, while if g_{XY} = 0.0, completely different sets of genes affect these traits. It is important to note that *heritability and genetic correlation are independent concepts*, so that highly heritable traits may share no genes (g_{XY} = 0.0) while weakly heritable traits can be influenced by exactly the same genes (g_{XY} = 1.0). Genetic correlations thus provide information about both *generalist* (g_{XY} > 0.0) *and specific* (g_{XY} < 1.0) *genes* (Plomin & Kovas, 2005:600-601).

Plomin and Kovas (2005:602, Table 4) report the genetic correlations between various aspects of language (including reading and writing) and find that, generally, these correlations are very high[131] (e.g., $g_{lexic,grammar} \approx 0.61$), suggesting that, for example, for vocabulary and grammar, the genetic factors overlap substantially (Plomin & Kovas, 2005:602). They conclude that, for spoken language,

> [a] two-factor model consisting of general language and articulation fit the data better than a single-factor model [but] the genetic correlation between these two latent factors was .64 [providing] strong evidence for the hypothesis of substantial genetic overlap among diverse aspects of language[132] (Plomin & Kovas, 2005:603);

the same broad conclusion seems to also hold for reading. These results provide evidence for both substantial genetic homogeneity of the language faculty and specificity of its components. More work is required for a better quantification of these overlaps and differences, but their impact on language impairment treatments and language evolution are potentially enormous.

130For details on its calculation see Plomin & Kovas, 2005:600-601.

131And, as for heritability, they change during development: for example, $g_{phonological\ awareness,\ grammar\ \&\ vocabulary}$ = 1.00 for 6 years old and 0.90 for 7 years olds (Plomin & Kovas, 2005:602).

132This conclusion applies mainly for normal abilities, as currently there are no such analyses for language-impaired populations (Plomin & Kovas, 2005:603).

The third question refers to co-morbidity, i.e., the co-occurrence of different disorders in the same individual (Plomin & Kovas, 2005:604). Stromswold (2001) reports the group heritabilities for language and IQ and concludes that

> [t]he similarity of h^2_g's for the populations of twins that included twins with low IQs and those that did not suggest that the heritability of language disorders is not merely the result of heritability of low cognitive function. [...] The genetic correlation between IQ and poor language achievement was less than .01, [indicating] that although genetic factors play a modest role in the phenotypic correlation between nonverbal IQ and language skills, *different genetic factors influence nonverbal IQ and language* (Stromswold, 2001:662, *italics* mine)

Moreover, given that $g_{\text{non-verbal delay, verbal delay}} \approx 0.36$, genes responsible for non-verbal delay and those responsible for low verbal scores in non-verbal delayed probands are mostly different (Stromswold, 2001:663). Plomin & Kovas (2005:604-607) report genetic co-morbidities for language, mathematics and reading and conclude that the genetic correlation between them are substantial (Plomin & Kovas, 2005:605). They also highlight the important point that double dissociations, usually taken to prove modularity (e.g. Pinker, 1995, 1997), occur even when the genetic correlations are high, because they follow a bivariate normal distribution: what is important to show is that the frequency of double dissociations is greater than expected from this distribution (Plomin & Kovas, 2005:606), and conclude that "[...] genetic correlations are not 1.0, which means that there are specialist as well as generalist genes [...] [h]owever, what is interesting [...] is the great extent to which genes are generalists" (Plomin & Kovas, 2005:607). They offer a three-levels explanation for the existence of these generalist genes (Plomin & Kovas, 2005:607-613): *DNA/gene* (mainly, pleiotropy), *brain* (they seem to favor a complex network between genes, mechanisms and traits – see their Figure 5, Model 3, page 611) and *mind* (involving generic processes like working memory and the *g* factor).

3.1.7. Conclusions: genes and the capacity for language

I am using "capacity for language" as a very general concept subtending our biological characteristics making us able to learn and use language. In this context, it is clear that this "capacity" is very much influenced by our genes, the strength and type of influence varying with the particular aspect under focus. It seems that the *many genes with small effects* model is the best explanation for the vast majority of language disorders as well as the normal range of variation in language abilities, and that some of these genes are *generalists* (influencing

various aspects of language or even more) while some are *specialists* (influencing just some of its aspects). There might exist a limited number of subtending factors relevant to language, like *phonological short-term memory*, *acoustic processing*, or even *the g factor*. It also seems that many language disabilities represent, in fact, not qualitatively different entities but simply *the tail end* of the language abilities distribution. Single-gene disorders, like the famed heterozygous *FOXP2*-determined SLI are rare and special and do not shed much light on the more common forms of language pathology, nor on its evolutionary history. It is very much like trying to understand the evolutionary (cultural) history of internal combustion engines by studying single-point catastrophic effects of the fuel pipe ruptures – certainly relevant but much too limited.

These genetic influences on the linguistic faculty seem to suggest an *accretionary model for language evolution* (e.g., Pinker & Jackendoff, 2005:218; Corballis, 2004[133]; Parker, 2006a, b), in which many alleles with small (i.e. *continuous as opposed catastrophic*) effects, appearing at different times and in different contexts, some becoming fixed while some representing polymorphisms[134], helped build the modern language faculty. This clearly militates *against any catastrophic, single-mutation model* of language evolution (e.g., Crow, 2000; Lanyon, 2006) and *pro the continuing evolution of this capacity in modern humans*.

3.2. The Correlations between the distribution of languages and genes

Are genetic and linguistic *diversities* correlated in any meaningful way? Is there a connection between the two and if so, what are its causes and what methods can be used to study it? In order to answer to these questions, I will first briefly summarize what is known about linguistic diversity, and then move on to address the important problem of establishing links with modern human genetic diversity.

133I do not necessarily agree with his overall conclusions, especially the human-specific mutation *FOXP2* as the "*most recent event* in a sequence of genetic changes that honed vocal articulation to the point that speech could become fully autonomous [...]" (Corballis, 2004:548, *italics* mine).

134Either because not yet fixed or because of disruptive selection.

3.2.1. Linguistic diversity: patterns and explanations

The Ethnologue (Gordon, 2005) reports 6,912 languages spoken worldwide, with an average of 828,105 and a median of 7,000 speakers per language. These languages are distributed in 94 language families[135] (Gordon, 2005), 6 of which (Afro-Asiatic, Austronesian, Indo-European, Niger-Congo, Sino-Tibetan and Trans-New Guinea[136]) account for 84.75% of speakers and 64.87% of languages (Table 4).

Language family	*Languages*		*Speakers*			
	Number	*Percent*	*Number*	*Percent*	*Mean*	*Median*
Indo-European	430	6.22%	2,562,896,428	44.78%	5,960,224	150,000
Sino-Tibetan	399	5.77%	1,275,531,921	22.28%	3,196,822	18,686
Niger-Congo	1,495	21.63%	358,091,103	6.26%	239,526	26,000
Afro-Asiatic	353	5.11%	339,478,607	5.93%	961,696	20,151
Austronesian	1,246	18.03%	311,740,132	5.45%	250,193	3,384

Table 4: The 5 major language families in terms of number of speakers.
Adapted from Gordon (2005), Trans-New Guinea not included.

The remaining 15.25% of the world's spoken languages are distributed in 89 language families and 4 non-genetic groups (language isolates, mixed languages, creoles and unclassified). These 89 language families are small and include such controversial items as Australian. There are 36 language isolates accounting for only 1.18% of the world population, 19 mixed languages (0.01% speakers), 82 creoles (0.5% speakers) and 43 unclassified languages (0.01% speakers) (Gordon, 2005). The important difference between the average and median number of speakers points to a very interesting fact: 347 (5%) languages have more than 1,000,000 speakers and together account for 93.88% of the world population, while the remaining 95% of the languages are spoken by only 6% of the population (Figures 21 and 22). Most languages are spoken by tens of thousands of speakers but the bulk of world's population speaks one of the very few languages with more than a hundred million speakers. This skewed distribution requires an explanation (Nettle, 1998, 1999a, 1999b; Diamond, 1997, 1998; Diamond & Bellwood, 2003; Ostler, 2005; Cavalli-Sforza, Menozzi & Piazza, 1994; Bellwood & Renfrew, 2002).

135The difficulties associated with establishing language families are notorious and, thus, any count must be taken with a grain of salt (Trask, 1996; Campbell, 2004; Lass, 1997).
136This is especially controversial, defined mostly by exclusion.

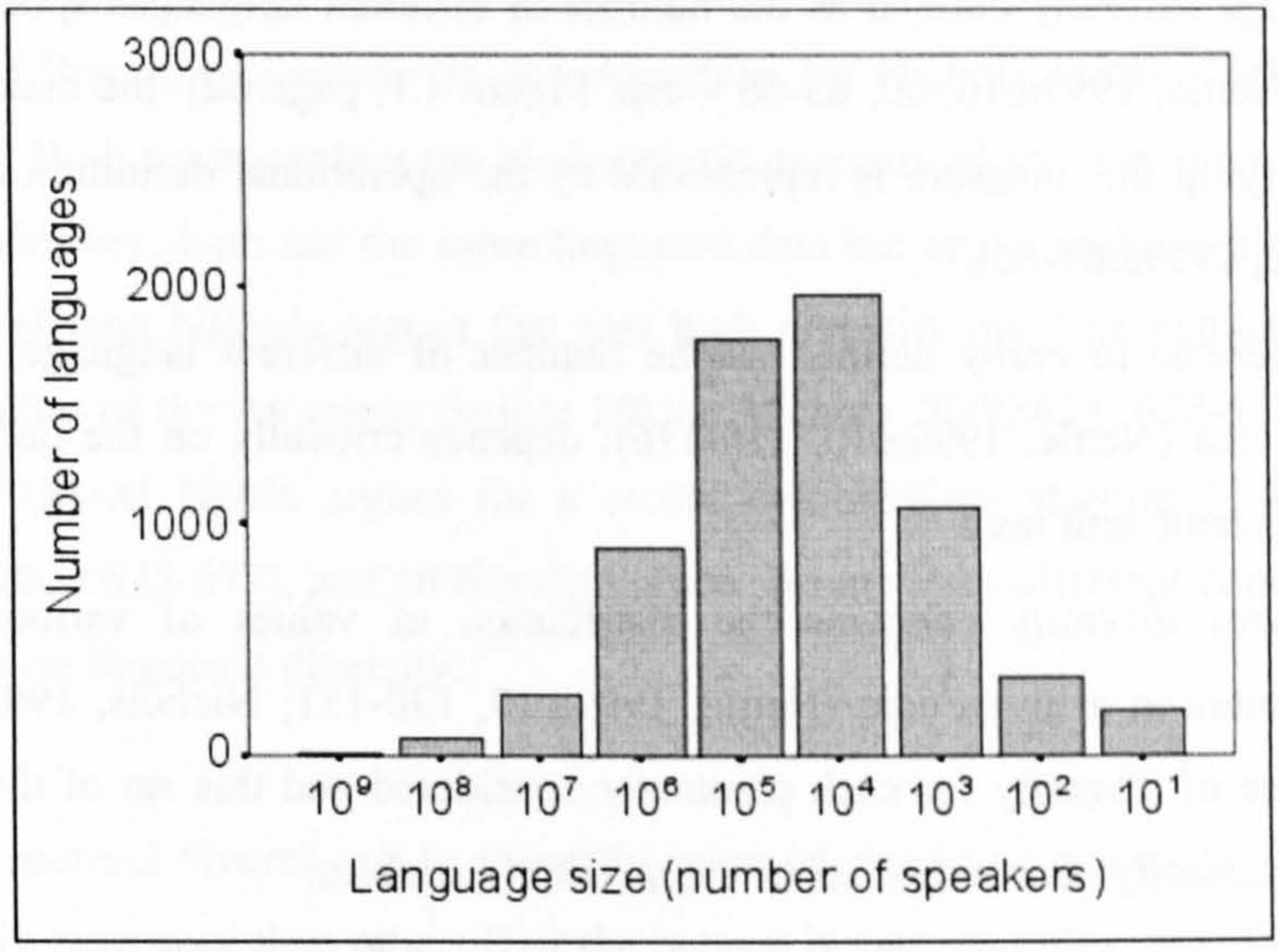

Figure 21: The number of languages of a given size.

Horizontal axis represents the language's size given as the number of speakers. Vertical axis represents the number of such languages. Drawn from data in Gordon (2005).

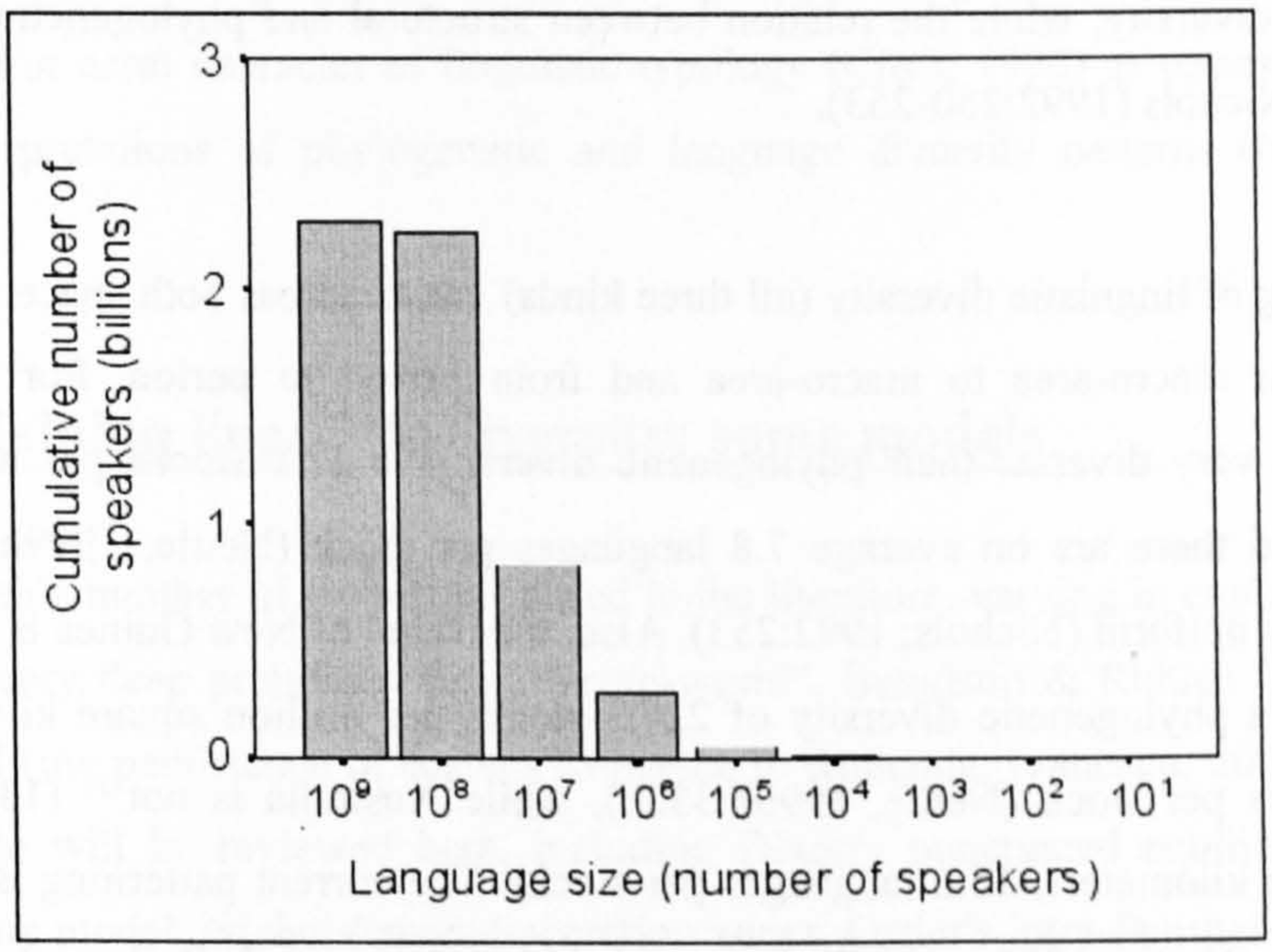

Figure 22: The cumulative number of speakers for languages of a given size.

Horizontal axis represents the language's size given as the number of speakers. Vertical axis represents the cumulative number of speakers. Drawn from data in Gordon (2005).

In his 1999 book, *Linguistic Diversity*, Daniel Nettle tackles this pattern of global language diversity and distinguishes three types (Nettle, 1999a:10):

- *language diversity* defined as the number of different languages spoken in a given area (Nettle, 1999a:10, 60, 63-66 – esp. Figure 4.1, page 62): the main difficulty in quantifying this measure is represented by the operational definition of a language (Nettle, 1999a:63-66);
- *phylogenetic diversity* defined as the number of different language lineages in a given area (Nettle, 1999a:10, 115-116): depends critically on the definition of the phylogenetic unit used[137];
- *structural diversity* concerns the distribution of values of various typological parameters in a given area (Nettle, 1999a:10, 130-131; Nichols, 1992): there is a measure of diversity for each parameter considered and this set of diversities does not necessarily constitute a set of independent variables.

There is no a priori correlation between these three types of diversity and any such statistically significant correlation found (globally or in a given area) needs an explanation. For example (Nettle, 1999a:10), Central Africa is high in language diversity but very low on phylogenetic diversity, while the relation between structural and phylogenetic diversities is discussed in Nichols (1992:250-253).

The patterning of linguistic diversity (all three kinds) varies across both space and time: it is different from macro-area to macro-area and from period to period. For example, the Americas are very diverse: their phylogenetic diversity is 27.2 stocks per million square kilometers and there are on average 7.8 languages per stock (Nettle, 1999b:3326), while Europe is very uniform (Nichols, 1992:253). Also, the island of New Guinea is exceptionally diverse, with a phylogenetic diversity of 227.3 stocks per million square kilometers, with 41.1 languages per stock (Nettle, 1999b:3326), while Australia is not[138] (13.0 stocks per million square kilometers, 15.6 languages per stock). The current patterning is the result of past processes (Nettle, 1999a, 1999b; Diamond, 1998; Diamond & Bellwood, 2003; Ostler, 2005; Bellwood & Renfrew, 2002), implying that the spatial and temporal aspects of linguistic diversity are not independent.

137Usually, Johanna Nichol's (1992) *stock*, roughly equivalent to the linguistic family (Nichols, 1992:24-26; Nettle, 1999a:116).

138The issue of Australia is very controversial: see Dixon (1997).

An example of conflicting models linking time and space is provided by the analysis and explanation of linguistic diversity in the Americas by Nichols (1992, 2000) and Nettle (1999b, 2000). Both try to explain the phylogenetic diversity of the Americas in light of the continent's prehistory, both use the same linguistic data but arrive at diametrically opposing conclusions. Johanna Nichols argues that this high diversity must be explained by a very early colonization of the Americas (before 20kya; Nichols, 2000:654, 658-661), in multiple waves, while Daniel Nettle argues for a recent colonization, starting 13-14kya (Nettle, 1999b:3328; 2000:675-677), and all this difference springs from different conceptions of the effects of time on linguistic diversity.

Concerning structural diversity, it is generally acknowledged that some typological features are more easily borrowed than others (but the pattern is very complex; see, for example, the papers in Aikhenvald & Dixon (Eds.), 2001 and especially Curnow, 2001), while Johanna Nichol's (1992) proposal concerning their different temporal stability is more controversial. In discussions of correlations between linguistic diversity and genetic diversity, structural diversity was not thoroughly considered to date, probably because of the perceived functional and/or areal character of linguistic typology (Croft, 1990) as opposed to directly historical interpretations of phylogenetic and language diversity patterns (Nettle, 1999a; Nichols, 1992).

3.2.2. Explaining linguistic diversity: some models

There is a sizable number of models proposed in the literature, varying in explanatory power from global, very deep prehistory (e.g., "proto-world", Bengtson & Ruhlen, 1994; Ruhlen, 1994) to local (the persistence of eastern Romance in Romania, Ivănescu, 2000), but only a limited sample will be reviewed here, including Dixon's punctuated equilibrium, Nettle's socio-economic model, Nichols' spread-accretion zones, Ostler's intra-familial language shift and the various forms of the language/farming co-dispersal hypothesis (Diamond, Renfrew, Bellwood, Cavalli-Sforza, etc).

The model proposed by **Dixon**, especially is his 1997 book *The rise and fall of languages* (Dixon, 1997), is inspired from Niles Eldredge and Stephen Jay Gould's punctuated equilibrium model of biological evolution (Eldredge & Gould, 1972). In biology, after a first

bout of controversy exacerbated by media exaggerations and misrepresentations, it seems that it represents not a mechanism but a result of biological evolution, as reflected in the very sparse fossil record (Skelton, 1993; Dawkins, 1990b, 1982: esp. 101-105; West-Eberhard, 2003). Dixon's *punctuated equilibrium* model of language change involves two tightly inter-related aspects of the same process: *equilibrium* and *punctuation*. As he explains equilibrium:

> In a given geographical area there would have been a number of political groups, of similar size and organization, with no one group having undue prestige over the others. Each would have spoken its own language or dialect. They would have constituted a long-term linguistic area, with languages existing in a state of relative equilibrium. Nothing is ever stasis – there would be ebbs and flows, changes and shifting around, but in a relatively minor ways (Dixon, 1997:3)

and the complementary process of punctuation:

> Then the equilibrium would be punctuated, and drastic changes would occur. [...] These punctuations to the state of equilibrium are likely to trigger dramatic changes within languages and between languages. They give rise to expansion and split of peoples and of languages. It is during a period of punctuation – which will be brief in comparison with the eras of equilibrium that precede and follow – that the family tree model applies (Dixon, 1997:3-4).

In Dixon's conception, the temporally dominant mode of language change is given by equilibrium states, where the prime process is represented by linguistic convergence/areal linguistics, while punctuations tend to be sudden, acute events rupturing the equilibrium states and allowing the "standard" linguistic families to develop. The causes of punctuations are multiple, including "[...] natural causes such as drought or flooding; or to the invention of a new tool or weapon; or the development of agriculture; or of boats, with movement into new territories; or to the development of secular or religious imperialism" (Dixon, 1997:3).

This model was primarily inspired by the author's experience with the Australian languages (Dixon, 1997, 2001) and his attempts at reconciling the "family tree" model of Indo-European, Afro-Asiatic and other such cases with the linguistic situation in other parts of the world, dominated by small linguistic families, many isolates and strong areal effects (Australia, New Guinean highlands, etc.). It has profound consequences on the understanding of current linguistic diversity, on one hand, arguing that the normal mode is that of equilibrium and large, well-structured families are recent, abnormal phenomena, bound to fade into a state of equilibrium, but also on the conceptualization of past linguistic phenomena, including the nature of protolanguages and the various macro-family claims. For example, he forcefully argues that protolanguages were *not* unitary languages which

expanded and split, but the results of long equilibrium states, more akin to linguistic areas which suffered expansions and differentiation (Dixon, 1997:97-99). He gives the example of Indo-European and Uralic as a complex linguistic process involving an initial stage of equilibrium between the protolanguages/proto-linguistic areas, followed by expansions and splits (Dixon, 1997:100, Figure 7.1, p. 101 – Figure 23 below).

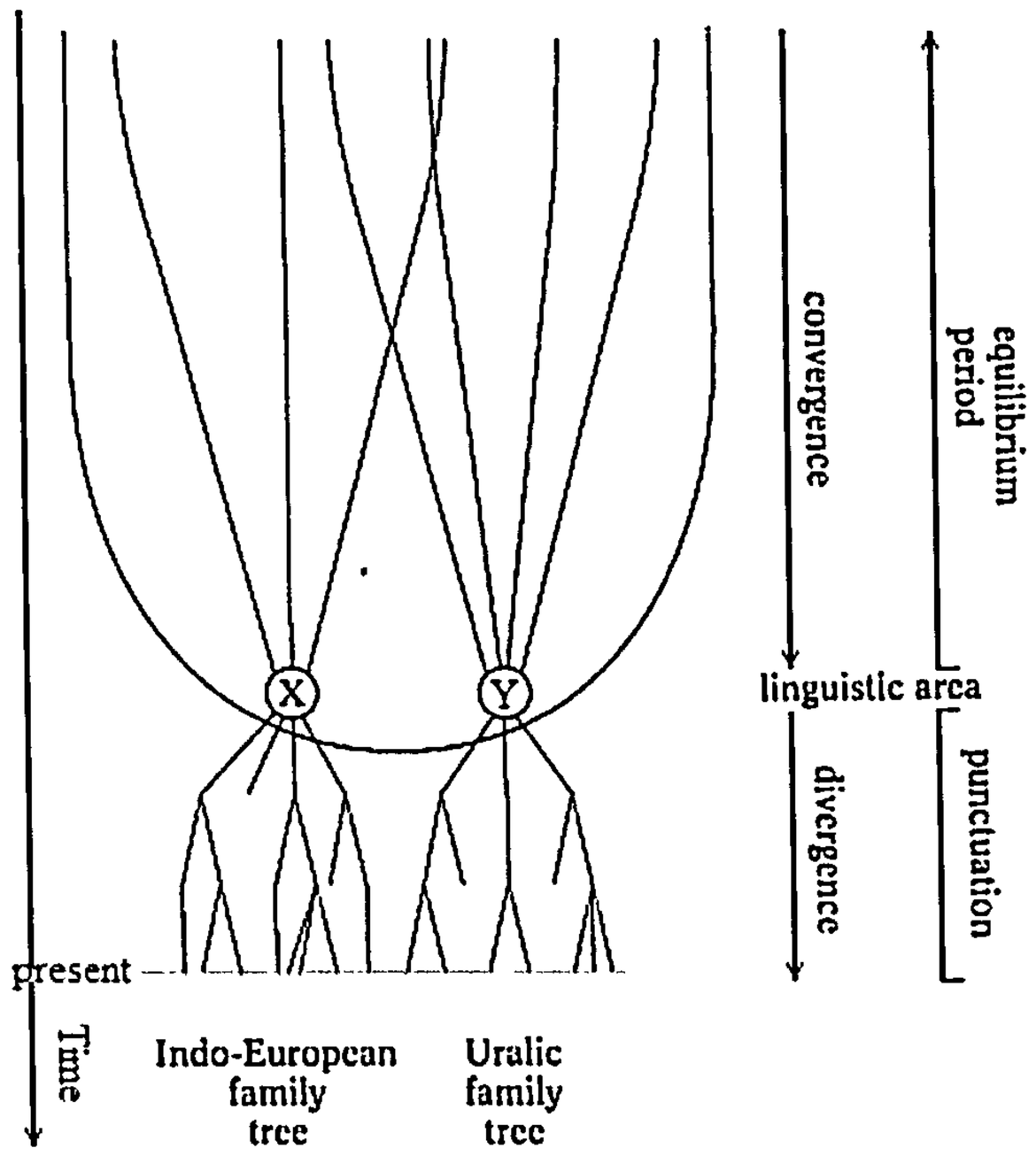

Figure 23: Dixon's example of equilibrium and punctuation for Indo-European and Uralic.

X and Y represent proto-languages (proto-linguistic areas) and the beginning of the two punctuation events are not necessarily simultaneous. Adapted from Dixon (1997), Figure 7.1, p. 101.

The apparent stability of the equilibrium periods is not a simple, static equilibrium but a dynamic one, and if we were to zoom in on such a period we would see a complex film of language contact, convergence, language shifts and differentiations, but on a local scale. Thus, the difference between equilibria and punctuations is one of degree and not of kind. In my opinion, this model is fundamentally different from its source of inspiration in biology, where the long periods of stasis are static equilibria: this fundamental difference is due to horizontal and diagonal transmission of language as opposed to vertical[139], across species.

139But see Section 2.2.4.

However, Dixon's own application of his model to explaining the evolution of language is unconvincing and superficial (Dixon, 1997:63-66), and he seems to favor a sort of naïve catastrophic theory for the emergence of language (Dixon, 1997:63). Nevertheless, what is very attractive in this model is its profound resemblance and compatibility with meta-population models of human evolution (Section 2.2.8). Overall, Dixon's punctuated equilibrium seems very promising, even if it needs refining and testing[140].

Daniel **Nettle**, in a series of papers (Nettle, 1998, 1999b, 1999c, 2000), but especially in his 1999 book (Nettle, 1999a), develops a model of linguistic diversity which I will call the "*socio-economic model*". Its basic tenets are represented by a non-monotonic relationship between time and diversity (Nettle, 1999b, 2000), a non-constant rate of change (Nettle, 1999c, 2000) and a socio-economic network constraining language sharing (Nettle, 1999a, 1998). He identifies two main types of social bonds: *primary* (very strong, enduring, formed early in life, multivalent and generalized; Nettle 1999a:67, 1998:359) and *secondary* (specifically functional, associated with greater social distance and temporally circumscribed; Nettle 1999a:67, 1998:359-360). Language spreads across the social networks formed by primary bonds, while secondary bonds are more often associated with relations between ethnolinguistic groups (Nettle, 1999a:67). Given the assumption that "the spread of a language is rooted in an economic system. [...] Choosing a particular dialect gives access to particular networks of cooperation and exchange that have material as well as social costs and benefits" (Nettle, 1999a:69), it is to be expected that linguistic (and specifically, *languages*) diversity will correlate with these socio-economic networks, and because in pre-industrial societies this is inseparable from ecology, it follows that "ecological risk is the most important influence on human social networks [and languages diversity]" (p. 70).

Ecological risk is defined as "the amount of variation which people face in their food supply over time" (Nettle, 1998:362) and in pre-industrial societies it determines the intensity and spatial spread of the primary social network, allowing the populations to cope with it (Nettle, 1999a:79-81). By increasing the spatial extent of these socio-economic networks, two benefits arise: access to a higher diversity of micro-ecologies, allowing a *spatial averaging* of risk, and also a *numeric averaging* simply because different households will probably be

140It must be noted that Dixon (1997) was received with some hostility by Australian linguists, but further work is needed in order to test his theory.

affected to various degrees (Nettle, 1999a:81). On a selected sample of languages[141] and using the *mean growing season* (MGS)[142] as a proxy for ecological risk, Nettle formulates and tests two hypotheses concerning languages diversity: increased ecological risk decreases the languages density and increases the number of speakers per language (Nettle, 1999a:83). Using a multiple regression approach (Tabachnick & Fidell, 2001:111-176), the dataset supports the two hypothesis (Nettle, 1999a:84-93; 1998:365-368). It seems, thus, that for pre-industrial agricultural societies, ecological risk is a good predictor of languages diversity.

Concerning the temporal dimension, Nettle is very much influenced by Dixon's punctuated equilibrium (see above). He identifies a very long *Palaeolithic equilibrium*, attributable to pre-agricultural hunter-gatherers (Nettle, 1999a:99, 100-103), following their expansion across the Old World from Africa and later into the Americas. He uses the historical pre-European contact Australian aborigines as a model for Palaeolithic hunter-gatherers (Nettle, 1999a:101), a highly controversial assumption, given the peculiarities of the Australian continent (Diamond, 1998; Mithen, 2003). Also, a fission-fusion model is assumed (Nettle, 1999a:102), implying that the number of languages has increased roughly in a linear fashion with population size, giving an overall estimate of 1600-9000 languages for the late Palaeolithic. The *Neolithic punctuation* represents the expansion of a limited number of language families through a farming/languages co-dispersal mechanism (see below), and the replacement of many other hunter-gatherer languages (Nettle, 1999a:103-105). This was followed by the *Neolithic aftershock* (Nettle, 1999a:105-108), including such dispersals and replacements on a massive scale as the European colonial period, followed by the *industrial punctuation* (Nettle, 1999a:108-112), continuing today, and further reducing the world's linguistic diversity. A specific but important critique concerns the *non-homogeneity of climate* (and thus, ecology and geography) during the Palaeolithic (which comprises major glaciations and intergacials; Wilson, Drury & Chapman, 2000; Mithen, 2003; Jobling, Hurles & Taylor-Smith, 2004), which determined in turn huge demographic alterations. Moreover, these changes have a sharp regional pattern[143], making any global inferences very difficult.

141The study is restricted to the tropics, small countries and countries with very heterogeneous ecological risk were excluded (Nettle, 1999a:82-83). Even if some criteria used to select this data set can be criticized (e.g., using the country as the areal unit, exclusion of heterogeneous countries, etc.), the results seem valid as a first approximation.

142A given month "is included in the growing season if the average daily temperature is more than 6°C and the total precipitation in millimeters is more than twice the average temperature in centigrade" (Nettle, 1999a:82). The MGS is simply the country's average of growing season across its weather stations.

143For example, the LGM (Wilson, Drury & Chapman, 2000) affected differently Europe (ice-cap

Thus, it must be concluded that the notion of a Palaeolithic equilibrium is not well supported by the available data.

The best description of Johanna **Nichols'** model is provided by her 1992 book (Nichols, 1992). Here, she identifies two main categories of areas from a linguistic diversity point of view (Nichols, 1992:13-16). The first is represented by *spread zones* (Nichols, 1992:16-21), which can be briefly characterized as large areas of low diversity, due to easy large-scale demographic events. Their characteristics are:

> (1) Little genetic [i.e., phylogenetic] diversity, a property that can be quantified as low genetic density (the ratio of genetic stocks to million square miles of area [...]). Most spread zones have genetic densities that are about half that of their continents. Often, a single language family dominates the spread zone.
>
> (2) Low structural diversity.
>
> (3) The language families present in the spread zones are shallow [i.e., recent].
>
> (4) Rapid spread of languages or language families and consequent language succession.
>
> (5) Classic dialect-geographical area with innovating center and conservative periphery. The center is a center of cultural, political, and/or economic influence. The center may shift as political and economic fortunes shift.
>
> (6) No net long-term increase in diversity. A spread zone is a long-lasting phenomenon, but it preserves little evidence of its history.
>
> (7) The spreading language serves as a lingua franca for the entire area or a large part of it (Nichols, 1992:16-17),

and some classic examples are western Europe, Australia, North America (Nichols, 1992:17) and central Asia. This idea of spread zones is very popular and usually associated with important migrations/conquests over very large areas, due to the specific geography (no major barriers, east-west dominant orientation) and ecology (homogeneous) and has a lot of explanatory power. It was further refined by the proponents of farming/co-dispersal hypothesis, not only in relation to this specific phenomenon (Diamond, 1998).

The second type is represented by *residual* (Nichols, 1992) or *accretion* (Nichols, 1997) *zones*, which can be briefly defined as those areas where languages tend to accumulate over long periods of time (Nichols, 1992:21-23). Their characteristics are:

> (1) High genetic [i.e., phylogenetic] diversity, significantly higher than the overall density of the host continent, often an order of magnitude higher
>
> (2) High structural diversity

advance), Australia (dryness) and South-East Asia (very little disturbance) (Mithen, 2003).

> (3) The language families, or at least a good number of them, are deep [i.e., old]
>
> (4) No appreciable spread of languages or families. No language succession
>
> (5) No clear center of innovation. Despite this (and despite the high genetic and structural diversity), there are usually some clear areal features
>
> (6) Accretion of languages and long-term net increase in diversity. Language isolates and isolate families are likely to be found in residual zones
>
> (7) No lingua franca [...] for the entire area; local bilingualism or multilingualism is the main means of inter-ethnic communication (Nichols, 1992:21),

and typical examples include the Caucasus, northern Australia and California (Nichols, 1992:21). One can visualize spread and accretion[144] zones as complementary results of the same process, whereby successive waves of incoming languages advance and become established throughout a spread zone and the old languages (including the previous dominant ones) survive in the adjacent peripheral accretion zones. Accretion zones usually pose problems for large-scale demographic movements and provide means for almost complete self-sufficiency of small communities (e.g. mountains) (Nichols, 1992:21-22), allowing thus the accumulation over time of linguistic diversity, in all its forms. Moreover, given this and the time depths involved, it is conceivable that areal features arose and spread, forming linguistic areas.

This classification has been criticized on many grounds (e.g., Campbell, 2002) and can be summarized as:

- problems with *classifying zones* (Campbell's 2002:56 "missassignment" problem): some zones are classified as spread when in fact they to not match the appropriate criteria;
- *language representativeness* (Campbell's 2002:56 "language representatives" and "area double-dipping" problems): this is a general problem in linguistic diversity studies and not specific to Nichols' approach, concerning the languages chosen to represent a given area and/or phylogenetic unit;
- *distinguishing* between accretion and spread zones (Campbell, 2002:56): the classification is subjective and depends on non-linguistic indices (historical records, etc.).

Another important critique concerns the non-uniformity of process: there is no a priori reason to expect that successive spreads behaved in the same way, given that different

144I will use henceforth the term of accretion zone, as it seems both more frequently used in recent literature and more suggestive of the processes involved.

constraints controlled their initiation and subsequent development. For example, changes in climate, technology or subsistence mechanisms affect the resulting linguistic pattern, so that what was a spread zone in earlier times could become impenetrable, and thus a (potential) accretion zone. Nichols briefly discusses this possibility (Nichols, 1992:20-21), exemplifying with western Europe, but the problem is more pervasive and potentially important. Nevertheless, if we accept the concepts of spread and accretion zones as fuzzy, aimed at initial exploration and orientation of more specific inquiries, then they are useful, at least as rough approximations and descriptions of a much more complex reality (*pace* Campbell, 2002:16-17).

Another look at the present linguistic diversity is offered by Nicholas **Ostler**'s 2005 book (Ostler, 2005), where a historical approach to languages is taken. He describes the history of a specific set of languages, reconstructed from written records, and tries to understand the historical conditions allowing some of them to dominate the current linguistic map of the world. This set comprises Sumerian, Akkadian, Phoenician, Aramaic, Arabic, Turkic, Persian, Egyptian, Chinese, Sanskrit, Greek, Celtic, Latin, Germanic, Slavic, Nahuatl, Quechua, Chibcha, Guaraní, Mapudungun, Spanish, Portuguese, French, Russian and English, and besides the wealth of data provided by this history, very important generalizations about linguistic diversity are given. One of them concerns the strategies a language has to become spoken by a large population (Ostler, 2005:19): *organic growth* (a language community which stays united while constantly increasing in size through demographic growth; "*the Farmer's Approach*") and *merger & acquisition* (increase in the number of speakers through language shift, due to migration, diffusion and infiltration – a combination of migration an diffusion; "*the Hunter's Way*"). He demolishes the power-based explanations of language shift (military conquest, political domination, religious activities) and provides a series of convincing examples which, even if apparently seeming to support such explanations (English, Latin, Arabic), are in fact better explained by socio-economic processes of the merger & acquisition type (Ostler, 2005:20-22). In the same vein, he identifies (mass) migration coupled with demographic explosion as the most important factor in language spread (Ostler, 2005:534-535), while trade and religion have played minor roles (Ostler, 2005:536-537). Concerning *prestige*, he also criticizes the received wisdom:

> [a] prestige language, in general, is any foreign language that is learned for cultural advantage. Sumerian, Akkadian, Chinese, Sanskrit, Greek, Latin, Arabic, Turkish, Persian, Italian, French, German and English have all been such

> languages in their time. But the time will not last forever. To be a prestige language, its native speakers – or the written records they have left – must somehow impress, and so attract imitators. This impact will depend on the cultural development of the recipients, as well as the merits of the originals. As potential recipients grow in wealth, knowledge and self-confidence, and begin to distinguish themselves, the attraction of a foreign model will shrink (Ostler, 2005:552).

He emphasizes *adult learners* as the actors of language shift, making thus the problem of *language learnability* by such agents essential for language spread (Ostler, 2005:552-556). The case of adult second language learning is the commonest situation when languages spread and might impose some interesting constraints relevant for the possible succession of languages (Ostler, 2005:553): "it might cause the learners to come up with a *new version of the language*, influenced by their old speech" (p. 553, *italics* mine) as, for example, the English spoken in India, or,

> more radically, the constraint may act as a *major block on the learners ever gaining effective command of the new language*. An example of this might be seen in the widespread failure of English Language Teaching (ELT) in Japan for several decades after the Second World War, despite Herculean efforts on all sides to give the next generation competence in this new skill (Ostler, 2005:553-554, *italics* mine).

A good example is represented by the spread (and failure to do so) of Arabic:

> [i]t settled permanently only in the territories that had previously spoken an Afro-Asiatic language, i.e. one that was structurally close to Arabic itself. First of all, Arabic took over the Aramaic-speaking world [...] [,where it] could have replaced Aramaic almost word for word. It then overran quickly, and subsequently pervaded, the countries of North Africa, whose vernacular was Egyptian [...] and Berber, although in these cases the spread was far slower, and – at least in the case of Berber – is by no means complete (Ostler, 2005:554),

while in Spain and Persia, even if these were early centers of Islamic and Arabic scholarship (Ostler, 2005:554; Hourani, 2002), the language did not replace the previous Indo-European languages (Ostler, 2005:554). Other such examples include Greek in western Asia and Egypt, Mongolian in central/western Asia and Europe and Latin in Gaul (structurally similar) as opposed to British Celtic (structurally divergent)[145] (Ostler, 2005:555-556). And he concludes that:

> Overall, it seems that – despite the received wisdom of linguists over two centuries and more – there may be circumstances in which the very essence of a language, its structure, can play a role in its viability. Languages, we suggest, are more easily learnt by a new population, and hence spread more easily, when they

145The same type of explanation was sometimes put forward also for the rapid replacement of Dacian by vulgar Latin.

are structurally similar to the old language of that population (Ostler, 2005:556).

It seems, thus, that language shift is principally a matter of adult second language learners and depends crucially on the *structural similarity* (*learnability*) of the target language relative to their first language. This does not affect in any way the principle that all languages are equal in all relevant linguistic aspects (expressivity, learnability by children etc.) but does impose powerful constraints on the dynamics of linguistic diversity, as *this process would tend to preserve areal structural features over long periods of time and across multiple language shifts*.

These various explanations of linguistic diversity must, probably, be combined in order to obtain a globally (as well as locally) acceptable model, but this is still far from being achieved.

3.2.3. The language/farming co-dispersal hypothesis

The modern distribution of linguistic diversity is very unbalanced, with a minority of language families accounting for a majority of speakers (Section 3.2.1). One popular explanation for this is represented by the generic proposal that some language families were spread together with agriculture, replacing the languages of the indigenous hunter-gatherers in the process. This type of theories is best championed by Jared Diamond, Peter Bellwood and Colin Renfrew, and this section will review their theories and the (non-genetic) data supporting or attempting to falsify them.

Agriculture, and its synonym, farming, are defined as "[t]he science and art of cultivating the soil; including the allied pursuits of gathering in the crops and rearing live stock; tillage, husbandry, farming (in the widest sense)" (OED, "agriculture") but, besides this broad sense, the term is also used to mean specifically "[...] the intensive farming of crops and animals in fields, as distinct from a less intensive management of individual plants (horticulture) and the breeding of animals (pastoralism)" (Jobling, Hurles & Tyler-Smith, 2004:300). The transition from previous hunting and gathering economies to farming was a gradual process, prompted by the climatic instabilities at the end of the LGM.

The global climatic oscillations have complex causes, but some very important forcing

factors during the Quaternary seem to be represented by the Milanković[146] cycles (Wilson, Drury & Chapman, 2000:61-65; Mithen, 2003:11)[147]. The shape of Earth's orbit around the Sun changes between more and less circular (*eccentricity* between 0.005 and 0.058, mean 0.028) in cycles of approximately 95 and 400ky, combining into an ~100ky cycle. When the orbit is more elliptical (high eccentricity), the seasonality is increased in one hemisphere and decreased in the other. The tilt of Earth's axis of rotation relative to the plane of its orbit around the sun (*obliquity*) also varies with an amplitude of about 2.6° (21.8° - 24.4°) with a ~41ky periodicity, and the greater the tilt, the greater the seasonality. The Earth's axis of rotation describes a full circle during a 27ky cycle (*precession*) and it also causes variations in seasonality. Another cycle of 105ky concerns the *precession* of the Earth's orbit around the sun, so that the perihelion occurs at different dates around the year, also impacting seasonality (Figure 24). The connection between these cycles and the onset of Ice Ages on Earth is provided by the particular current configuration of the continents (Wilson, Drury & Chapman, 2000:59-61), due to continental drift (Marshak, 2005). The Northern Hemisphere contains a large area of landmasses at high latitudes, which are susceptible to supporting large ice caps[148] when climatic conditions became favorable: warm and wet winters (high snow fall) and cold summers (low ice melt). These climatic conditions are favored by specific configurations of the astronomic cycles, recurring with a periodicity of ~100ky.

146Named after the Serbian geophysicist Milutin Milanković (1879-1958).

147Even if there are some problems, this theory seems supported for the moment (Wilson, Drury & Chapman, 2000:82-112 for a discussion); the phenomena are extremely complex and more than one explanation must be envisaged (Wilson, Drury & Chapman, 2000:139-159).

148Which, in turn, through their increased albedo, determine a positive feedback, favoring their own expansion (Wilson, Drury & Chapman, 2000:59).

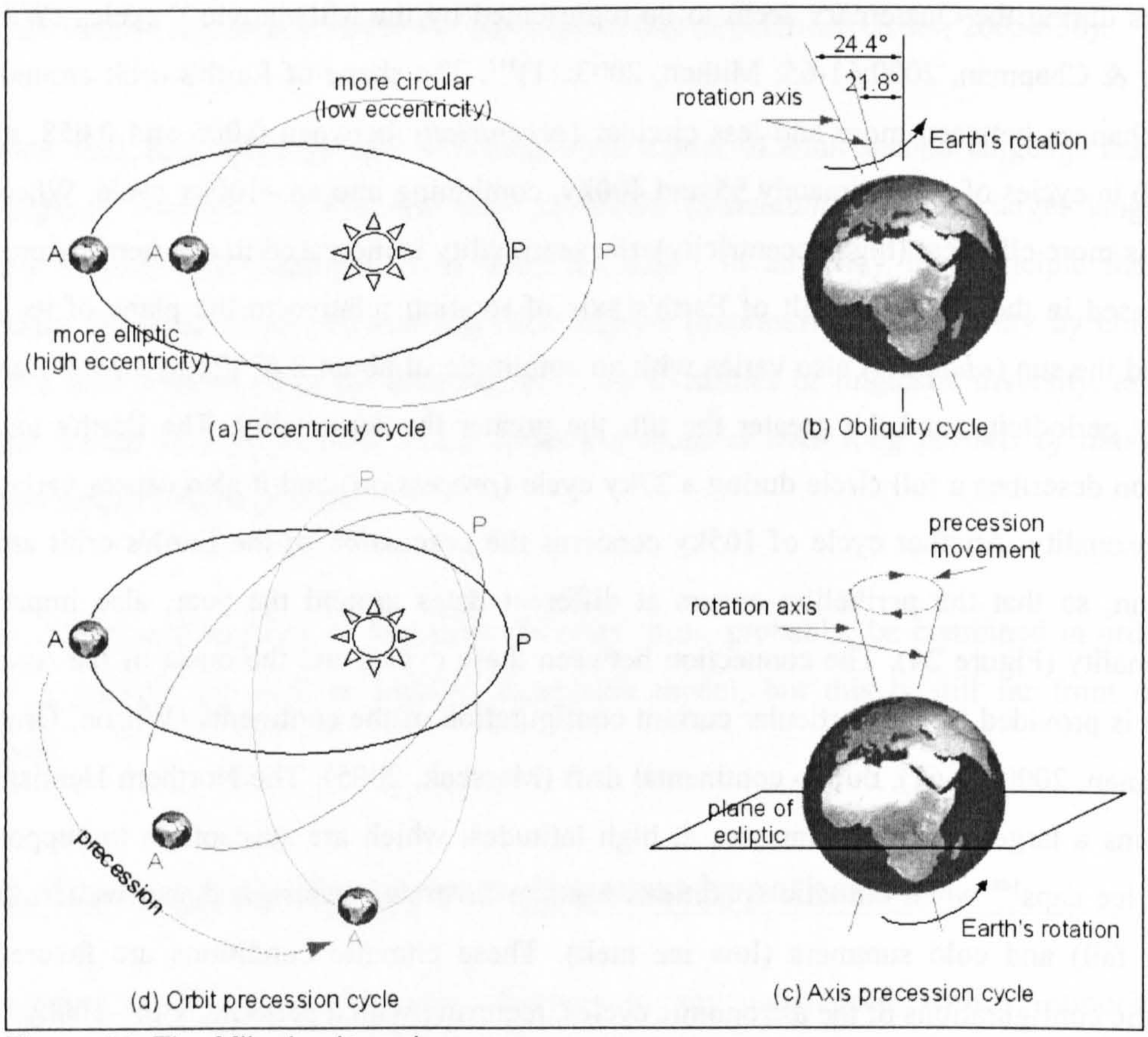

Figure 24: The Milanković cycles.

*(a) the **eccentricity** cycle, modifying the shape of the Earth's orbit (95 & 400ky), (b) the **obliquity** cycle, modifying the tilt of the Earth's rotation axis (41ky), (c) the **precession** of the Earth's rotations axis (27ky) and (d) the **precession** of the Earth's orbit (105ky). Adapted from Wilson, Drury & Chapman, 2000:63-64. See text for details.*

After the LGM, a period of global warming begun ~15kya, with a very unstable climate for the first ky and punctuated by the Younger Dryas event[149] ~12.8-11.6kya, an abrupt and short return to Ice Age conditions (Mithen, 2003:12; Wilson, Drury & Chapman, 2000), after which the climate stabilized at interglacial conditions (the Holocene). A representation of the climate fluctuations during the last 25ky is given in Figure 25, drawn using data from the NGRIP database (NGRIP, 2006), containing the $\delta^{18}O$ analysis of the NGRIP1, NGRIP2, GRIP and DYE-3 ice cores from Greenland (NGRIP, 2006). $\delta^{18}O$, or delta values of the ^{18}O

149Named after the flower mountain avens (*Dryas octopetala*), which flourished in Europe during this period (Mithen, 2003:113). Its causes are still debated and include a short interruption in the North Atlantic termohaline circulation (Wilson, Drury & Chapman, 2000:153-154).

oxygen isotope[150], refer to the proportion of ^{18}O and ^{16}O isotopes[151] in a sample, given with reference to a standard[152] and is measured in parts per thousands (‰, 'per mil'). Given that water containing the ^{18}O isotope ($H_2{}^{18}O$) evaporates slower and condenses easier than "normal" water ($H_2{}^{16}O$), the $\delta^{18}O$ varies with the average temperature, so that it represents a *proxy* for temperature (in ice cores, the lower $\delta^{18}O$, the lower the temperature) (Wilson, Drury & Chapman, 2000:72-75).

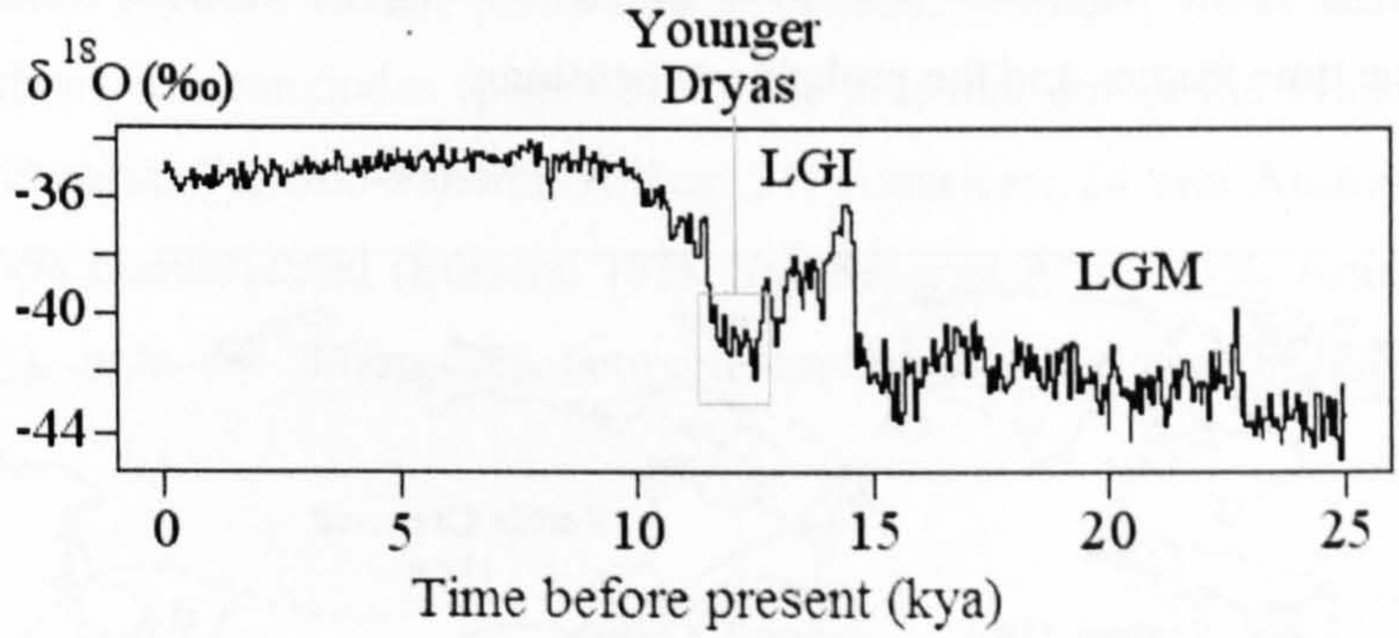

Figure 25: The climatic record of the last 25ky.

It shows the Last Glacial Maximum (LGM) ~20kya, the Last Glacial Interstadial (LGI) 12-14kya and the Younger Dryas event, 11.8-9.6kya. Drawn from NGRIP data (NGRIP, 2006).See text for details.

The end of the LGM and the ensuing global warming allowed the expansion of human populations both geographically (in areas previously uninhabitable: covered with ice sheets, deserts) and numerically (increase in the carrying capacity of many habitats due to warmer and wetter climate), but it also provoked disturbances of the rich seashore habitats due to sea level changes (Mithen, 2003; Diamond, 1998; Fagan, 2004). In this climatic context, the onset of agriculture was a very complex process (Diamond, 1998; Mithen, 2003; Bellwood & Renfrew, 2002; Fagan, 2004).

Diamond & Bellwood (2003:597), assert that, following the climatic stabilization after the Younger Dryas,

> [...] at different subsequent times between 8500 and 2500 B.C. [10.5-4.5kya], food production based on domestication of relatively few plant and animal species arose independently in at most nine homelands of agriculture and herding,

150Computed as $\delta^{18}O = 1000 * ((^{18}O/^{16}O)_{sample} - (^{18}O/^{16}O)_{standard}) / (^{18}O/^{16}O)_{standard}$ (Wilson, Drury & Chapman, 2000:72).

151The ^{16}O isotope is the most common (>99%), with ^{18}O accounting for most of the rest (^{17}O extremely rare).

152The Standard Marine Ocean Water (SMOW) (Wilson, Drury & Chapman, 2000:72).

scattered over all inhabited continents, except Australia[153] (Diamond & Bellwood (2003:597).

The actual number of areas of independent agricultural onset is highly contentious (Jobling, Hurles & Tyler-Smith, 2004:301), but at least the Fertile Crescent, China and Mesoamerica seem uncontroversially accepted (Jobling, Hurles & Tyler-Smith, 2004:301; Mithen, 2003). The map in Figure 26 represents the areas of agricultural innovation (both primary and secondary), the time frames and the probable expansions.

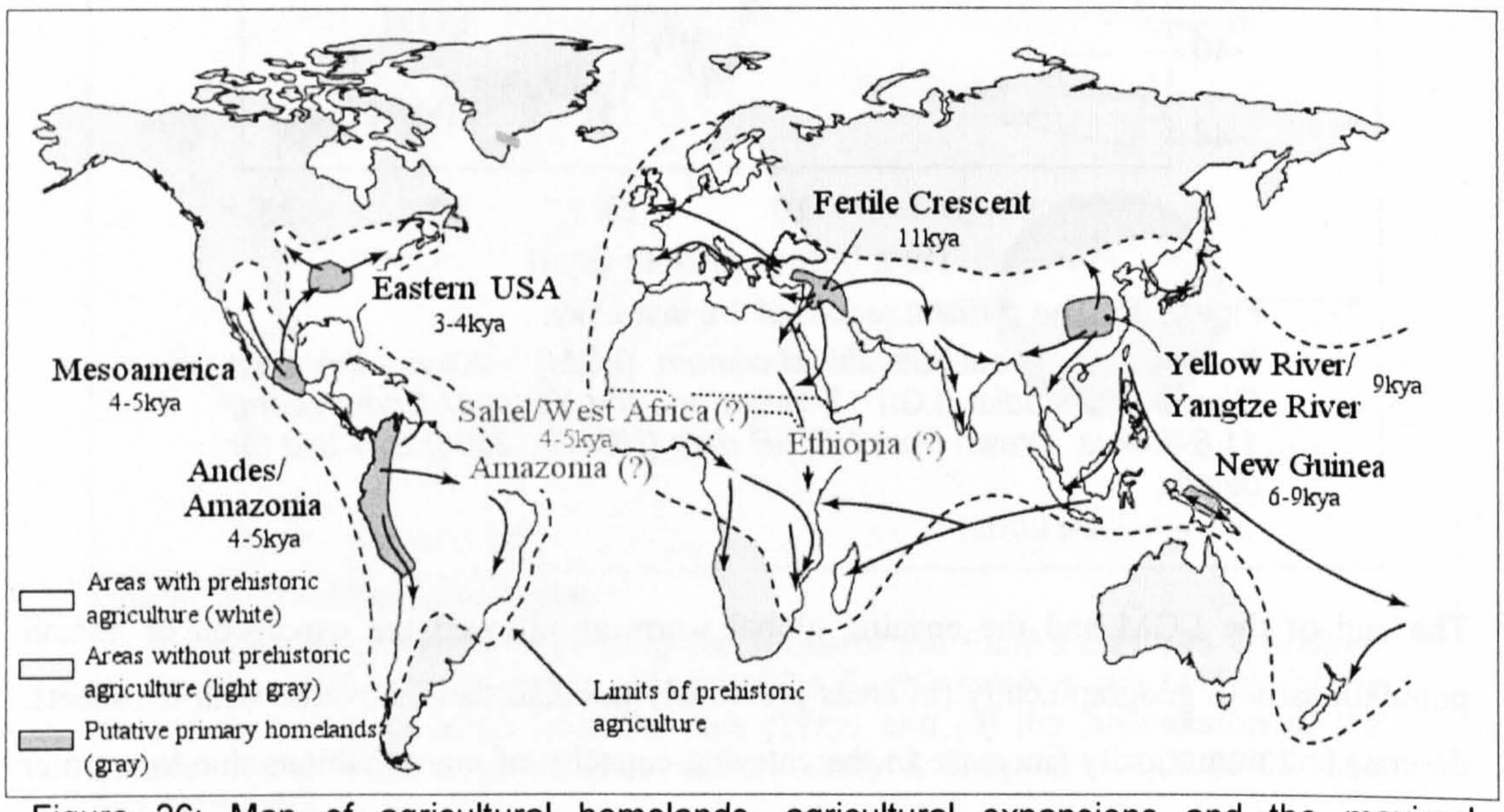

Figure 26: Map of agricultural homelands, agricultural expansions and the maximal prehistoric agricultural area.

Based on Diamond & Bellwood (2003:597, Fig. 1), Jobling, Hurles & Tyler-Smith (2004:302, Figure 10.2; 303: Figure 10.3) and Bellwood (2003:18, Figure 2.1). Gray areas represent primary homelands (Eastern USA, Mesoamerica, Andes/Amazonia, the Fertile Crescent, Yangtze/Yellow River Basins and New Guinea highlands) and the appropriate timeframes for the onset of agriculture. Gray writing represents possible but unconfirmed primary homelands (Amazonia, Sahel/West Africa, Ethiopia). White areas represent the maximal extension of prehistoric agriculture while light gray areas represent the areas without prehistoric agriculture. Arrows represent expansions. See text for details.

The main specific domesticates for each region are in Jobling, Hurles & Tyler-Smith (2004:303-305), Mithen (2003) and Diamond (1998); but see the papers in Bellwood & Renfrew (2002) for the controversies in each case. New domesticates could have been added

153The failure of Australia to develop agriculture and the geophysical and climatic factors responsible are detailed in Diamond, 1998.

in secondary areas and old ones dropped, especially because of local eco-climatic conditions. There is a set of necessary conditions for a plant or animal species to be domesticated, reducing very much the number of possible domesticates in each area (Jobling, Hurles & Tyler-Smith, 2004:305, Table 10.2; Diamond, 1998:157-175; Mithen, 2003). For example, Diamond (1998:160-161, Table 9.1) lists the 14 species of big herbivorous mammals domesticated during Prehistory (*the major 5*: sheep, goat, cow, pig and horse and *the minor 9*: arabian camel, bactrian camel, llama/alpaca, donkey, reindeer, water buffalo, yak, bali cattle and mithan) and concludes (page 162, Table 9.2) that out of the available candidate mammals[154] (Eurasia: 72, Sub-Saharan Africa: 51, Americas: 24 and Australia: 1), only a tiny percent was domesticated (Eurasia: 18%, Sub-Saharan Africa: 0%, Americas: 4% and Australia: 0%), with the differences between continents fully accountable by objective factors[155].

The actual causes of the transition to agriculture are contentious and it seems more appropriate to search for local explanations in the global context of the climatic instability and subsequent stabilization of interglacial conditions following the LGM[156] (Diamond, 1998, 2002; Mithen, 2003; Fagan, 2004; Jobling, Hurles & Tyler-Smith, 2004:303-304), but what seems to be a very powerful explanatory device is represented by the *irreversibility of farming* (Jobling, Hurles & Tyler-Smith, 2004:304; Diamond, 1998, 2002; Mithen, 2003; Fagan, 2004): farming can support larger population densities than hunter-gathering, so that once a population became *dependent* on farming, there is no turning back, insuring that farming will eventually displace hunting and gathering whenever environmental and demographic conditions allow.

The exceptions to this pattern, represented by communities of hunter-gatherers not adopting farming and not overrun by farmers, living in areas suitable for agriculture and in which agriculture could have expanded, provide an interesting cue to the origins of farming. Such

---

154These highly imbalanced figures are due to continent-specific factors (size, geography, history), including the differential impact of the late Pleistocene extinctions, partially attributable to humans (Diamond, 1998; Mithen, 2003).

155I.e., not connected in any way to the characteristics of their inhabitants, *contra* racist discourses attributing them to the "mental inferiority" of the natives (Diamond, 1998).

156For example, the history of agriculture in the Near East/Fertile Crescent proves to be very complex, involving the sheer luck of multiple wild ancestors of domesticated plants and animals living in close proximity in the same area, coupled with the climatic fluctuations of the Younger Dryas, forcing less efficient new subsistence patterns to emerge (Diamond, 1998:104-175; Mithen, 2003:20-96; Fagan, 2004; Jobling, Hurles & Tyler-Smith, 2004:301-305; Bar-Yosef, 2002).

communities are represented, for example, by the North-Western coastal Native Americans (Mithen, 2003:296-300) and the explanation involves the fact that in certain rich ecological environments[157], the hunter-gatherer lifestyle allowed high population densities and social structure, thus counterbalancing both the demographic pressures of neighboring farming communities and the need to shift to agriculture as a means to manage the ecological risk (Mithen, 2003; Diamond, 1998, 2002). This supports a view of the transition to farming as a means to deal with ecological risk, as a non-preferred strategy compared to hunting and gathering, and not as a "progressive" move "waiting" to be discovered. It seems, in fact, that primitive farming was initiated and forgotten many times, following the fluctuations of climate and preceded by what is called the *Broad-Spectrum Revolution*[158] (Cohen, 2002:41; Bar-Yosef, 2002:114; Mithen, 2003), pointing to the fact that people were *forced* to adopt farming by ecological/climatic factors and not that agriculture was hard-to-discover, hidden and requiring a sort of genius. In fact, hunter-gatherers were seemingly very much aware of the drawbacks of early farming lifestyles, including the increased disease burden, poor overall nutrition and social conflicts and inequality (Diamond, 1998; Mithen, 2003; Bellwood & Renfrew, 2002). Thus, the transition to farming was gradual, involving many early reversals, due to ecological and climatic forcing and depending on specific continental factors.

> Once domestication began to arise, the changes of plants and animals that followed automatically under domestication, and the competitive advantages that domestication conveyed upon the first farmers (despite their small stature and poor health), made the transition from hunter-gatherer lifestyle to food production *autocatalytic* – but the speed of that transition varied considerably among regions (Diamond, 2002:701, *italics* mine).

The transition to farming (the "*Neolithic revolution*": Diamond, 1998; Jobling, Hurles & Tyler-Smith, 2004; Mithen, 2003) arguably had a set of important consequences for the current distribution of human genetic and linguistic diversity (Jobling, Hurles & Tyler-Smith, 2004:305-306; Mithen, 2003; Diamond, 1998, 2002): higher population densities, increased population growth rates, malnutrition, epidemic infectious diseases and societal

---

157Intensive salmon fishing, in this case (Mithen, 2003:297-298).

158This is roughly equivalent to *Mesolithic* in the Old World and *Archaic* in the New World and refers to:

> "the increasingly intense utilization of the diverse resources of a small geographical area, including among other things an increased use of resources such as small game, riverine, coastal and lacustrine resources such as shellfish, and small seeded plants, often accompanied by increasing processing of foods (e.g., grindstones), storage, and (semi-) sedentary lifestyles" (Cohen, 2002:41).

changes. These changes, taken together, produced demographic, technological and military advantages for the farming compared to hunter-gatherer societies, which resulted in an overall replacement of the latter by the former. The details are very complex and regionally specific, but seem to have involved a combination of demographic replacement of hunter-gatherers (with various degrees of admixture) and cultural shifts of the hunter-gatherers themselves to farming. The papers in Bellwood & Renfrew (2002) and Sagart, Blench & Sanchez-Mazas (2005) and the discussions in Diamond (1998, 2002), Mithen (2003), Cavalli-Sforza, Menozzi & Piazza (1994) and Fagan (2004) address the complexities of this transition, its global patterns and local details.

The two extreme models for the expansion of agriculture are represented by *cultural* versus *demic diffusion*[159] (Jobling, Hurles & Tyler-Smith, 2004:300; Renfrew, 2002; Cavalli-Sforza, Menozzi & Piazza, 1994:102-103; Cavalli-Sforza, 2002:80-83, 108-111). In the cultural diffusion (acculturation) model, local populations of hunter-gatherers adopted farming from neighboring populations, allowing thus the spread of agriculture through cultural shift and without (or with negligible) population replacement, while the demic diffusion model argues that local hunter-gatherers did not shift to agriculture but were replaced by, or incorporated into, the incoming wave of farmers, numerically, technologically and militarily superior, allowing thus the spread of agriculture with the agriculturalists themselves (Jobling, Hurles & Tyler-Smith, 2004; Cavalli-Sforza, Menozzi & Piazza, 1994; Bellwood & Renfrew, 2002). The emerging consensus seems to be that these two models are complementary and that their relative importance depended on local demographic, geographical, ecological and cultural conditions (Mithen, 2003; Bellwood & Renfrew, 2002).

If acculturation is dominant, one would expect a decorrelation between culture and genes, in the sense that the cultural construct of farming spread while the people's genes (mostly) did not, but if the demic diffusion model is dominant, then we would expect at least a partial correlation between genes and cultures to be present, in the sense that they spread together, carried by the farmers themselves (Jobling, Hurles & Tyler-Smith, 2004; Cavalli-Sforza, Menozzi & Piazza, 1994; Bellwood & Renfrew, 2002; Mithen 2003). One of the consequences of a demic diffusion model with admixture is that the resulting genetic pattern is represented by a gradient radiating from the center of expansion, due to introgression of

159Or *wave of advance* (Cavalli-Sforza, Menozzi & Piazza, 1994:108-109).

local ("hunter-gatherer") genes into the "farming" gene pool. Moreover, if the acculturation process is strong enough in a given area, the local, previously hunter-gatherer populations adopting farming, would be able to induce a higher level of genetic introgression into the farming gene pool. This scenario, which seems very likely as a general description of the transition to agriculture, predicts that the correlation between the cultural trait (farming) and the genetic traits is not perfect and decreases with increasing geographic distance[160] from the center of expansion (Figure 27). The details of such models, their predictions and shortcomings are discussed, for example, in Cavalli-Sforza, Menozzi & Piazza (1994:3-157), Jobling, Hurles & Tyler-Smith (2004:300-301, 309-312), Renfrew (2002:10-14), Cavalli-Sforza (2002:80-87), Hurles (2002:300-302), Zvelebil (2002:379-386), Barbujani & Dupanloup (2002:421-422, 426-428) and Chikhi (2002).

Demic diffusion should produce genetic gradients, which, theoretically, can be detected in living populations (Cavalli-Sforza, Menozzi & Piazza, 1994; Jobling, Hurles & Tyler-Smith, 2004), but the fundamental assumption is that the expanding farmers and local hunter-gatherers are genetically distinct (Figure 27, population 4 at time 7). This, in turn, introduces the further complication that hunter-gatherer populations tend to be more genetically similar with decreasing geographical distance, thus increasing the apparent degree of admixture for populations closer to the center of expansion[161].

160Also taking into account the possible geographic, ecological and cultural barriers to farming.

161For proposed solution to such problems see, for example, Cavalli-Sforza, Menozzi & Piazza (1994), Chikhi (2002).

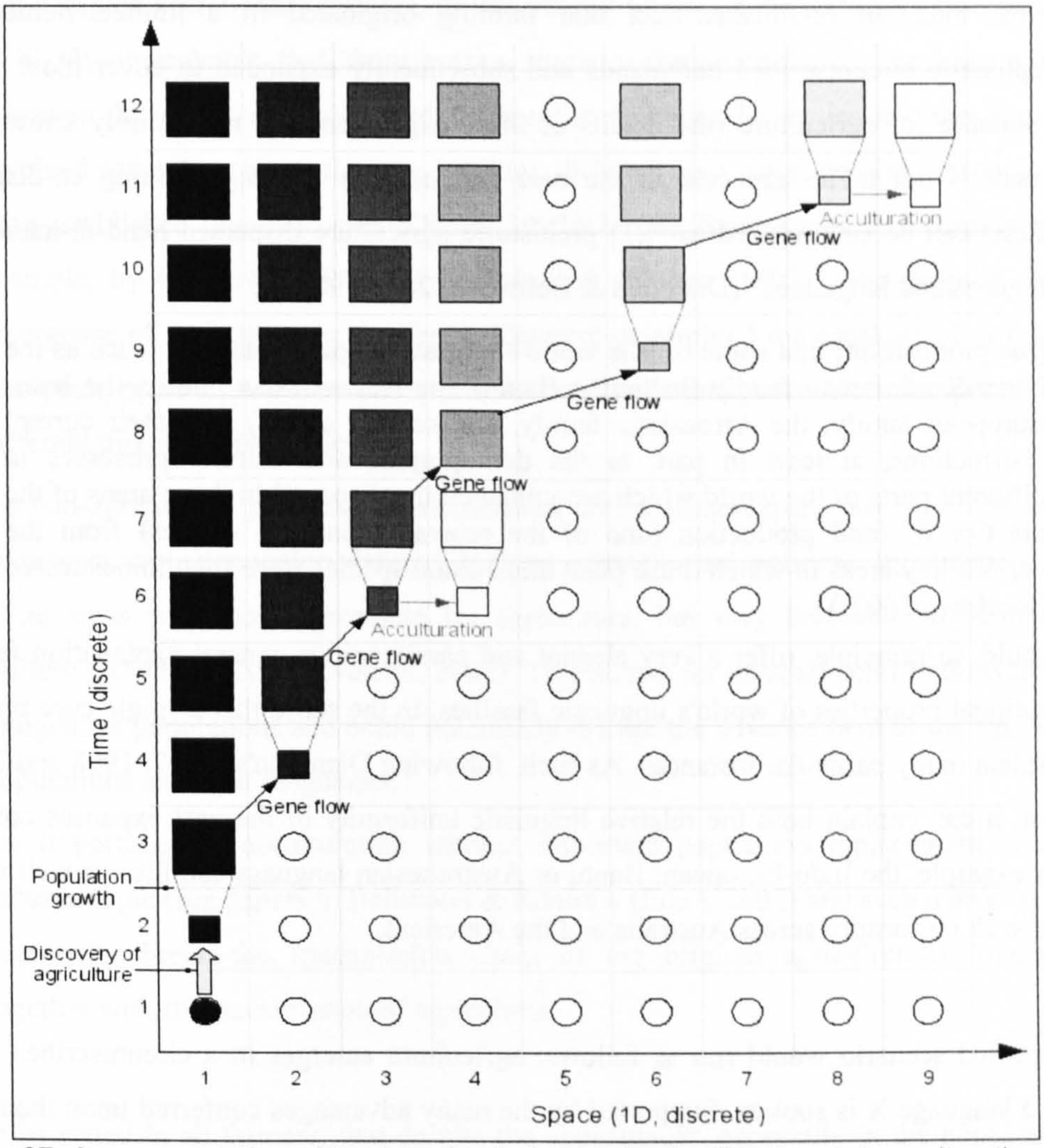

Figure 27: A representation of the interplay between demic diffusion and acculturation in the spread of farming.

Space is unidimensional (1D) and discrete (divided in regions, 1 to 9) and time is also discrete (1 to 12 periods). Circles represent hunter-gatherers while squares represent farming populations. The shades of gray (from black to white) represent genetic ancestry and admixture. At time 1, every population pursues a hunter and gathering lifestyle but at time 2 population 1 (represented as black) "discovers" agriculture and grows (time 3), so that at time 4 it spreads into the neighboring region 2 (gene flow, black diagonal arrows), populated by hunter-gatherers (white) and, following unequal admixture, generates a new farming population which traces its ancestry mostly to the original farmers and some to the local hunter-gatherers (represented as a slightly paler shade of gray). The process is replicated for region 3 at time 6, but now the hunter-gatherers living in region 4 shift to farming through acculturation (gray horizontal arrows), so that their population grows at time 7 and through equal admixture with incoming farmers produces the population at time 8 (equal admixture of population 3's gray and locals' white). Region 5 is unsuitable for agriculture, so that hunter-gatherers persist for a longer time, and farmers jump to region 6, and then to 8 (7 being also unsuitable). It must be remarked that if the indigenous farmers in region 4 at time 7 would have spread (gene flow) to the neighboring region 5, no gradient could have been detected. Adapted from Jobling, Hurles &Tyler-Smith, 2004:301, Figure 10.1.

It seems, thus, an established fact that farming originated in a limited number of geographically circumscribed homelands and subsequently expanded to cover most of the areas suitable for agriculture (the details of these expansions are not entirely known and agreement is not to be expected in the near future). The language/farming co-dispersal hypothesis can be summarized as "[...] prehistoric agriculture dispersed hand-in-hand with human genes and languages" (Diamond & Bellwood, 2003:598), or,

> [the proposition] that some of [the world's largest] language families (such as the Niger-Kordofanian family (including Bantu), the Austronesian family, the Indo-European family, the Afroasiatic family, and several others) owe their current distributions, at least in part, to the demographic and cultural processes in different parts of the world which accompanied the dispersal in those areas of the practice of food production (and of the relevant domestic species) from the various key areas in which those plant and animal species were first domesticated (Renfrew, 2002:3).

This could, in principle, offer a very elegant and parsimonious general explanation for the distributional properties of world's linguistic families, in the sense that a single core process can explain many particular instances. As such, following Diamond's (1997, 1998 and 2002) account, it can explain both the relative linguistic uniformity of the vast expanses covered by, for example, the Indo-European, Bantu or Austronesian language families as well as the lack of such uniformity across Australia and the Americas.

A simplified scenario would run as follows: agriculture emerges in a circumscribed area, where a language X is spoken. Supported by the many advantages conferred upon them, the speakers of language X start expanding in a wave-like manner, eventually performing frog-leaps between areas suitable to agriculture (Bandelt, Macaulay & Richards, 2002:104-105), and displacing or otherwise engulfing the indigenous populations. This demographic expansion is coupled with the expansion of language X, which, in the process, becomes differentiated both because of substratum influences from the languages of the original inhabitants and of divergence due to spatial separation, resulting, eventually, in a set of contiguous languages belonging to the same linguistic family. In Dixon's parlance, this would represent a punctuation event (Dixon, 1997:75, 77-78) *par excellence*, and the resulting linguistic family will allow a tree-based representation of these languages.

But the reality is much more complex and a sample of the possible intervening factors is:

- given that the emergence of agriculture was not an instantaneous "discovery" but a long

and complex accretionary process, driven mainly by geo-climatic factors (see above), it is highly improbable that there was a single unitary population speaking a single language X which expanded and differentiated. More likely is that the expansion involved several such populations, speaking different languages, possibly belonging to a long-established linguistic area (Dixon, 1997:97-102). Such a model is considered, for example, by Renfrew (1991) and Barbujani & Pilastro (1993), who propose that the emergence of agriculture in the Fertile Crescent determined the speakers of proto-Indo-European, proto-Elamo-Dravidian, proto-Afro-Asiatic and proto-Altaic to spread in four different directions (see below);

- the non-agricultural populations encountered during these expansions were not uniform in respect with population density, military strength and political organization, while some areas were not appropriate for agriculture, but very favorable to hunting and gathering (Mithen, 2003; Zvelebil, 2002). This allows for unequal contributions from the indigenous populations and could potentially disrupt the advancement of the agricultural populations and their languages;
- the importance of acculturation without important population replacement is a very debated topic (see papers in Bellwood & Renfrew (Eds.), 2002) and such a process could potentially break the transmission chain of the original agriculturalists' languages together with the transmission of agriculture.

It can be stated in all fairness, that despite the appearances, especially in the popularization press, the language/farming co-dispersal hypothesis is far from being universally accepted. Testimony to this controversy is the collection of papers in Bellwood & Renfrew (Eds., 2002), some of which espouse opposing stances on the same circumscribed issue. Concerning particularly the Indo-European case, one could consult, for example, Mallory (1991) and Sims-Williams (1998). It must be noted that in the case of the Polynesian branch of Austronesian, the language/farming co-dispersal hypothesis can be regarded as probably true, given that the expansion of the agricultural populations happened relatively recently in an uninhabited territory (Sagart, Blench & Sanchez-Mazas (Eds.), 2005; Diamond, 1998; Jobling, Hurles & Tyler-Smith, 2004:354-370).

One type of argument adduced in this debate, concerns the potential informativeness of the pattern of modern genetic diversity on past demographic events relevant for the current

linguistic diversity. This type of data is usually highly regarded (Diamond, 1998; Cavalli-Sforza, 2000; Cavalli-Sforza, Menozzi & Piazza, 1994) and applied to a diverse set of problems, ranging from the identity of the Etruscans (Vernesi *et al.*, 2004), through the Indo-European homeland and expansion in Europe (Cavalli-Sforza, Menozzi & Piazza, 1994:296-299), to the problem of language origins (Cavalli-Sforza, 2000). Following the terminology previously defined, these represent cases of spurious correlations between genetic and linguistic diversities and their assumptions, methods and results will be analyzed in the following section.

3.2.4. Spurious correlations between genetic and linguistic diversities

Population genetic techniques can be used to discover past demographic events, including migrations, admixture, expansions or bottlenecks and such events could shed light on linguistic phenomena concerning the distribution of various linguistic groupings (dialects, languages, sub-families, families or macro-families[162]).

There are a number of genetic methods (Jobling, Hurles & Tyler-Smith, 2004, esp. Chapters 5 and 6) which can be used. One is based on a set of neutral alleles: their frequencies are measured in as many populations as possible and a database results, on which different statistical techniques are applied in order to detect patterns of genetic diversity, including Principal Components Analysis, the detection of boundaries and computation of various genetic distance measures.

Principal Components Analysis (PCA) as applied to genetic data was made popular by the seminal work of Cavalli-Sforza and co-workers (Cavalli-Sforza, Menozzi & Piazza, 1994; Cavalli-Sforza, 2000; Cavalli-Sforza, Menozzi, Piazza & Mountain, 1988; Cavalli-Sforza, Menozzi, Piazza & Mountain, 1989; Ammerman & Cavalli-Sforza, 1984). PCA can be briefly described as a statistical method of compressing a large number of variables into a smaller number of *components* summarizing most of the variance in the original data[163]. These components are linear combinations of the original variables and are chosen to be orthogonal (independent) and to account for the maximum amount of variance in the data.

162The usage of this concept does not imply my agreeing with it (see below).

163The technique is very close to *Factor Analysis*: see Tabachnick & Fidell (2001:582-585) for common and divergent points.

The resulting components (PC1, PC2, ...) are sorted in decreasing order of how much variance in the original data they account for. In principle, there are as many PCs as variables in the original set, but there are conventional strategies for selecting the minimal number of PCs explaining most of the variance (Tabachnick & Fidell, 2001:582-652; Cavalli-Sforza, Menozzi & Piazza, 1994:39-42).

Results are usually represented as maps of the PCs: the *factor scores*[164] of each population on the considered PC (say, PC1) are represented on the map at the population's geographical location. Furthermore, different *interpolation techniques* (Fortin & Dale, 2005:159-170) are used to estimate the values of the locations between the sampled populations, so that a map representing continuous changes in factor scores (*gradients*) is produced (Cavalli-Sforza, Menozzi & Piazza, 1994:50-52; Jobling, Hurles & Tyler-Smith, 2004:170, 187-189) (Figure 28). But despite looking nice and compelling, the interpretation of such maps is difficult and controversial (Cavalli-Sforza, Menozzi & Piazza, 1994; MacEachern, 2000; Sims-Williams, 1998). The main problem is that, fundamentally, they represent interpolated values of summarized allele frequencies, which means, on the one hand, that the smoothing process used to produce the gradients can hide discontinuities, boundaries or even different local gradients, while, on the other hand, differences in allele frequencies can be due to many overlapping processes, including gene flow, population movement and natural selection (Jobling, Hurles & Tyler-Smith, 2004:125-150; Cavalli-Sforza, Menozzi & Piazza, 1994:52-59). The interpretation of these maps in terms of a single demographic process historically circumscribed (Cavalli-Sforza, Menozzi & Piazza, 1994), is, thus, at least hazardous and more akin to projective tests in psycho-diagnosis (Dumitraşcu, 2005; Meloy *et al.*, 1997) than to historical reconstruction (McMahon, 2004:9; MacEachern, 2000; Sims-Williams, 1998).

164The estimation of the values obtained by that population on the PC if someone would have managed to measure this directly (Tabachnick & Fidell, 2001:626-627)

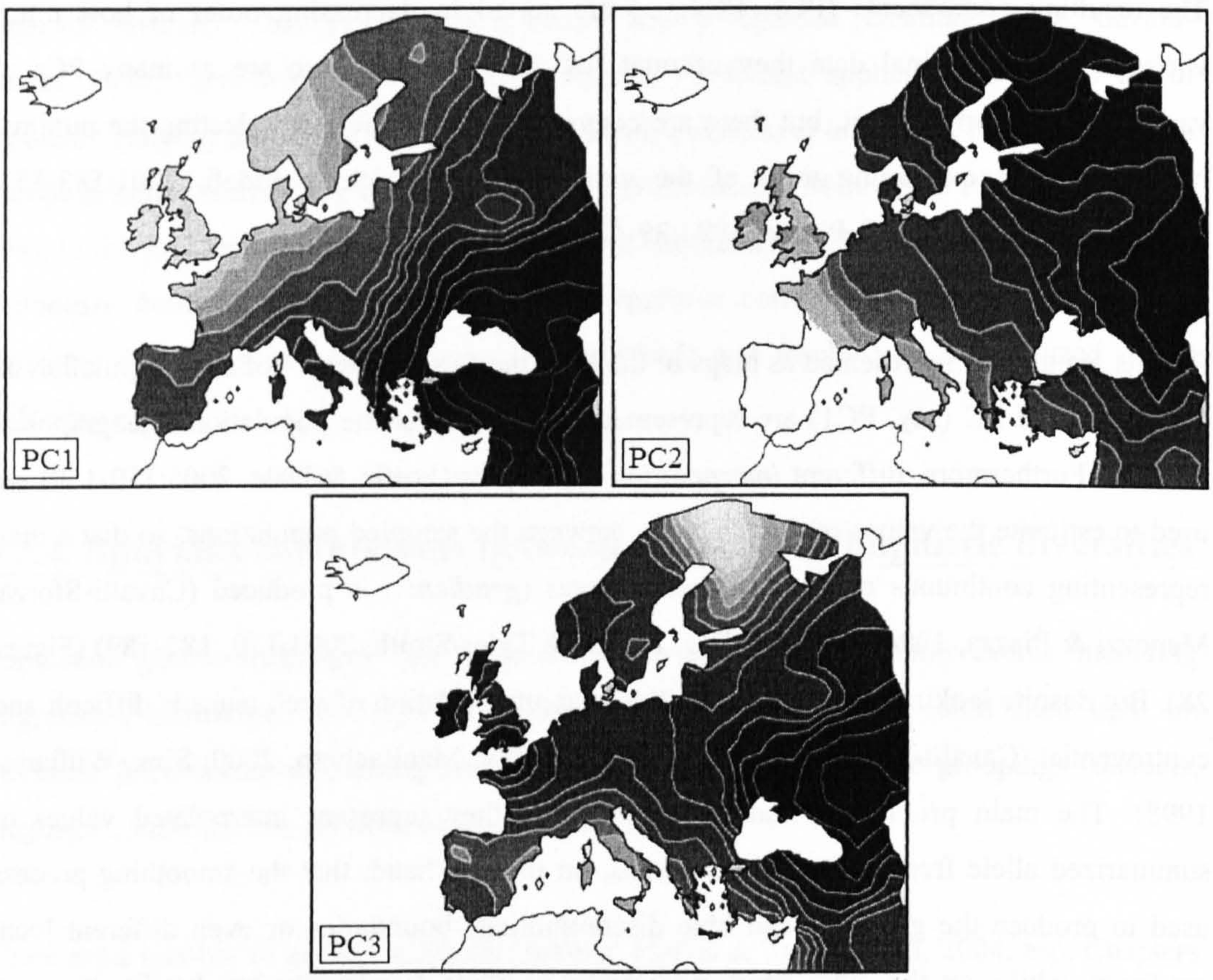

Figure 28: The first three principal components (PC1, PC2 & PC3) of 95 allele frequencies across Europe and the Near East.

PC1 accounts for 26% of the variation, PC2 for 20.6% and PC3 for 8.8%. PC1 has one extreme in NW Europe and the other in SW Asia, PC2 has a generic SW-NE direction while PC3 seems to radiate from the area NE of the Black Sea. PC1 is verbalized as showing an expansion from the Near East, PC2 as a concentric gradient radiating from the Iberian peninsula while PC3 as radiating from the Caspian steppe (Cavalli-Sforza, Menozzi & Piazza, 1994:291-293). The verbalizations of PC2 and PC3 are very subjective: PC2 can be seen as actually radiating from E-NE while PC3 as centered in the extreme North (Lapland). Their interpretation is even more prone to wishful thinking: PC1 is seen as representing the demic diffusion of early agriculturalists from the Near East spreading Indo-European languages, PC2 is interpreted as a climate-driven gradient while PC3 is taken to represent the Kurgan expansion, also carrying Indo-European languages. However, such gradients cannot be unequivocally associated with dates nor ethnic/linguistic labels, as they could be due to a multitude of phenomena, including migrations, natural selection, gene flow. Moreover, many successive such events, not necessarily following parallel geographic directions, are superimposed in a very complex palimpsest. For example, PC1 could be due to a putative neolithic expansion, to a Palaeolithic expansion, to post-LGM re-expansions, etc. or to any combination thereof. Note: the classic depictions of PC1, PC2 and PC3 in Cavalli-Sforza, Menozzi & Piazza (1994:292-293, Figs. 5.11.1-5.11.3) differ slightly from the ones reproduced here because of differences in the alleles used by Piazza et al. (1995). Adapted from Piazza et al. (1995, Fig. 1:5837, Fig. 2:5838 and Fig. 3:5839).

An alternative to the PCA/synthetic maps approach is the computation of various *genetic distances* between populations, usually related to F_{ST} or Nei's D (Jobling, Hurles & Tyler-Smith, 2004:166-170; Cavalli-Sforza, Menozzi & Piazza, 1994:29-30) and their interpretation in terms of proxy for historical divergence between populations. Such an approach also concerns the identification of *boundaries*, defined as "zones of abrupt genetic change" (Rosser *et al.*, 2000: 1532, Figure 6, p. 1538), but they seem rather uninformative, being mostly determined by geographical barriers (Rosser *et al.*, 2000:1537-1539; de Ceuninck *et al.*, 2000).

Another very popular, especially in the early literature, derivative is represented by the construction of *trees* out of such genetic distances between populations (Cavalli-Sforza, Menozzi & Piazza, 1994:30-39; Cavalli-Sforza, 2000:36-42) through such methods as Neighbor-Joining or UPGMA (Allman & Rhodes, 2004:171-198; Jobling, Hurles & Tyler-Smith, 2004:172-173; Cavalli-Sforza, Menozzi & Piazza, 1994:31-32). These methods take as input a matrix of distances (genetic distances, in our case) between any pair of entities (populations) and, through iterative clustering, return a tree such that entities separated by small distances tend to belong to lower-level sub-trees (Allman & Rhodes, 2004:180-198). The tree resulting from inputting genetic distances computed between populations is then interpreted as representing their genealogy, in the sense that diverging branches are regarded as population splits (Cavalli-Sforza, Menozzi & Piazza, 1994:38-39; Bateman *et al.*, 1990:7-8; Sims-Williams, 1998:520; Jobling, Hurles & Tyler-Smith, 2004:170).

But algorithms for tree construction from data matrices will always produce a tree, called a *phenogram*, no matter how inappropriate the original data is to be represented as a tree. There are methods available for assessing how well the tree "fits" the distance matrix or if the distance matrix can be reasonably well represented by a tree [e.g., *bootstrapping* (Jobling, Hurles & Tyler-Smith, 2004:175) or *treeness* (Cavalli-Sforza, Menozzi & Piazza, 1994:35-37, 57-59)], but the fact remains that, in interpreting a phenogram in genealogical terms, the implicit assumption is that the entities (populations) involved evolved separately after fission. As Peter Skelton warns:

> [i]t is important to reiterate that a phenogram is no more than a hierarchy of relative phenotypic similarity of a set of species. [...] It is *not intended as an accurate portrayal of phylogenetic relationships* [...] phenograms may still be misleading if treated literally as equivalent to phylogenies (Skelton, 1993:524,

italics mine).

There are two important points to note: first, he is referring to *different species*, where there is no gene flow which could potentially be a very important factor in the case of populations of the same species, making things even worse. Second, *phenograms do not necessarily imply measurements of the phenotype*: it simply refers to a method (*phenetics*) of collapsing data into an overall measure as opposed to *cladistic* methods, which treat each character state separately (Skelton, 1993:521-549; Bateman *et al.,* 1990:7). Thus, the simple employment of genetic data does not automatically shield a study from the shortcomings of phenetic methods, contra the unclear argumentation of Cavalli-Sforza, Menozzi & Piazza (1994:31), where they dismiss the concept of phenogram as applied to their methodology and replace it with the ambiguous "tree":

> [a]nother recently introduced term, *phenogram*, which is usually synonymous with dendrogram[165], is a misnomer when it refers to data on genotypes, such as those we employ. Should the trees we use be called *genograms*? *Tree* seems accurate and short, and if necessary it can be specified by the attribute "phylogenetic" (Cavalli-Sforza, Menozzi & Piazza, 1994:31, *italics* in original).

Moreover, they imply (again) that their phenetic trees represent phylogenies[166], while, in fact, they do not. As Skelton very clearly shows, phenetic methods can be applied to molecular genetic data without needing to coin new names for it (Skelton, 1993:550-554) and stand in sharp contrast to cladistic methods applied to the same type of data (Skelton, 1993:554-556), making thus the entire argumentation in Cavalli-Sforza, Menozzi & Piazza (1994) irrelevant[167]: their methodology, even if applied to genetic data, remains phenetic and the trees resulting from its application do not necessarily carry genealogical information about the concerned populations.

Another fundamental problem inherent in this approach is its assumption that trees are an adequate description of the history of human populations. Cavalli-Sforza, Menozzi & Piazza (1994:57-59) try to address the issue of admixture but treat it as the exception rather than the rule (MacEachern, 2000; Sims-Williams, 1998). Trees are not good descriptors of the history

---

165Defined in the previous paragraph on the same page as referring to the purpose of "formal [goals] – the processing of information with purely descriptive aims" (Cavalli-Sforza, Menozzi & Piazza, 1994:31).

166Defined as an "evolutionary tree" (Skelton, 1993:512).

167Historically, this argumentation must be understood as a reaction to the acid comments in Bateman *et al.* (1990, especially page 7 and footnote 8) and comments to this paper (Current Anthropology 31:13-24), but, unfortunately for Cavalli-Sforza and co-workers, the suggested change in name does not solve their deepest problems.

of populations belonging to the same species (Chapter 2), as populations rarely split and cease any form of contact for long periods of time[168] and admixture is the rule rather than the exception (MacEachern, 2000).

Another approach is to obtain information for non-recombinant genetic systems (Jobling, Hurles & Tyler-Smith, 2004:38-42): mitochondrial DNA (mtDNA) and the non-recombining part of the Y chromosome (NRY). They are much simpler to interpret historically than autosomal loci because they follow only the maternal and, respectively, the paternal lineage, while any autosomal locus reflects both histories at the same time. Also, by applying molecular dating techniques, it becomes possible to place mutational events in time (Barbujani & Dupanloup, 2002; Jobling, Hurles & Tyler-Smith, 2004:173-174, 177-183; Underhill, 2002:67). The information extractable from such genetic systems is potentially enormous, as proven, for example, by such studies as Tambets *et al.* (2002), Underhill (2002), Rosser *et al.* (2000), di Giacomo *et al.* (2004) or Poloni *et al.* (1997), but this type of genetic loci have intrinsic limitations, too, which are very often neglected when trying to interpret the discovered genetic patterns in a demographic, historic or linguistic context. These types of pitfall were thoroughly analyzed in connection with mtDNA as applied to the modern human origins debate (Chapter 2) and also apply in this context. It is never too often to repeat that, especially in the case of non-recombining loci, the reconstructed history is *not* the history of the population, but just the history of the transmitting sex (the female line for mtDNA and the male line for NRY), which, as important as it may be, represents a very partial point of view (Underhill, 2002:67). Moreover, mtDNA and the NRY tend to disagree more often than they agree (Jobling, Hurles & Tyler-Smith, 2004:314-323; Bandelt, Macaulay & Richards, 2002:463; McMahon, 2004:6), highlighting once more the importance of caution when generalizing inferences drawn from a single locus.

3.2.4.1. Some critiques of the language-genes studies

Interdisciplinarity is a highly prized endeavor in the currently fragmented scientific landscape (Bellwood & Renfrew, 2002:xiii; Cavalli-Sforza, Menozzi & Piazza, 1994:99-102, 372-373; Jobling, Hurles & Tyler-Smith, 2004:307-309) and was enormously successful in certain areas of human knowledge (cognitive science, human evolution, bio-chemistry, etc.).

168This is a general statement with possible exceptions, like the Polynesian islands or Australia after the LGM.

But good interdisciplinarity has very stringent requirements and goes far beyond the simplistic and superficial imports across disciplines, demanding a thorough understanding of the fundamental paradigms and concepts involved, and of the (often hidden) assumptions and limitations of the results in each discipline. Therefore, works trying to connect across the board, say, quantum mechanics and human consciousness (e.g., Penrose, 1989), using more than simplistic notions from psychology and neuro-sciences, cannot be qualified as good interdisciplinarity but as distortions of a very complex reality. Unfortunately, this critique seems to fully apply to some approaches trying to combine linguistics and human genetics into a unitary scientific field.

3.2.4.2. Superficial and incorrect usage of linguistic classifications

Probably the best illustration of the frustration felt by most linguists when faced with the inappropriate usage of linguistic classifications in the genes-languages literature (Bolnick *et al.*, 2004; Sims-Williams, 1998; Bateman *et al.*, 1990 and comments) is provided by Robert Dixon, cited *in extenso* below:

> Specialists in related disciplines take great interest in the family tree diagrams put forward by linguists. Archaeologists, geneticists and anthropologists *like to be given a clear-cut linguistic hypothesis*, about where and when a proto-language was spoken and exactly how it split and spread. They *happily accept any family tree that is produced*, without stopping to ask whether it is soundly based, and whether it is accepted by the majority of linguists. The excesses of Greenberg and the 'Nostraticists' have thus received acceptance outside linguistics itself. [...] When linguists tell archaeologists and geneticists that such and such a putative family tree is without scientific basis, the response is 'give us another family tree to replace it then.' If the linguist answers that the family tree model may not be applicable for the groups of languages in question – that it is a matter of typological similarity and linguistic area – the non-linguists may turn away with a shrug (and *will probably continue using the unjustified family tree*, just because they consider they need something like this, to tie their archaeological and genetic theories to) (Dixon, 1997:43-44, *italics* mine).

But is this critique justified?

In his 2000 book, "Genes, Peoples and Languages"[169], summarizing and popularizing his life-long approach to genes-languages interactions, Luca Luigi Cavalli-Sforza (Cavalli-Sforza, 2000) has an entire chapter dedicated to languages (Chapter 5, "Genes and

169The same critiques can be applied to the more technical Cavalli-Sforza, Menozzi & Piazza (1994), but these assumptions/arguments are not so visible.

Languages", pp.133-172), in which he argues that the linguistic classification used in all his earlier work (e.g. Cavalli-Sforza, Menozzi & Piazza, 1994, 1989) is sound, against all its critics: "The classification of families by Merritt Ruhlen (a student of Greenberg's) appears to me to be satisfactory for comparing genetic and linguistic evolutions [...]" (Cavalli-Sforza, 2000:139).

Ruhlen's linguistic classification (Ruhlen, 1987[170]) to which Cavalli-Sforza refers (Cavalli-Sforza, 2000:135, 139; Cavalli-Sforza, Menozzi & Piazza, 1994:22,-23, 96-98) is based on earlier classifications by Joseph Greenberg[171] (1963a, 1987) and recognizes 17 linguistic families: Khoisan, Niger-Kordofanian, Nilo-Saharan, Afro-Asiatic, Caucasian (later split into North and South/Kartvelian; Gordon, 2005), Indo-European, Uralic-Yukaghir, Altaic, Chukchi-Kamchatkan, Eskimo-Aleut, Elamo-Dravidian, Sino-Tibetan, Austric, Indo-Pacific, Australian, Na-Dene and Amerind. Their geographic distribution is given in Figure 29 (adapted from Cavalli-Sforza, Menozzi & Piazza, 1994:97, Fig. 2.6.1). But "[w]hile about half of these are well-established language families, the other half are speculative entities based mainly on geographical, anthropological, and plausible-guess criteria" (Nettle & Harriss, 2003:332), probably the most contentious being: Khoisan and Nilo-Saharan (Campbell, 1999:211-212 and references therein), Altaic (Campbell, 1999:204:210), Austric, Indo-Pacific (Dixon, 1997:34-35), Australian (Dixon, 1997:87-93; Dixon, 2001; Dench, 2001) and especially Amerind (Dixon, 1997:34-35; Bateman *et al.*, 1990; Bolnick *et al.*, 2004; McMahon, 2004:5; Sims-Williams, 1998:506, 520; Matisoff, 1990). The main arguments against these "linguistic families" are that the data used are not reliable or are even plainly wrong (Campbell, 1999; Dixon, 1997), that the methodology employed is not appropriate [e.g. Greenberg's "mass/multilateral comparison" (Greenberg, 1954, 1987) consisting basically in searching for lexical similarities between many languages; see Matisoff, 1990; McMahon & McMahon, 2005] and that they reflect areal phenomena and not genetic inheritance (Aikhenvald & Dixon (Eds.), 2001). The general opinion in the linguistic literature seems to be that they lack linguistic reality. Moreover, some even question apparently well-established families like Niger-Congo (Dixon, 1997:32-35; Campbell, 1999:212; Dimmendaal, 2001) or Afro-Asiatic (Campbell, 1999:210-215), but their status seems much safer for the moment.

170I will use the newer Ruhlen (1991).
171For an overview of his life and carrier, see his obituary by William Croft in *Language* 77:815-830.

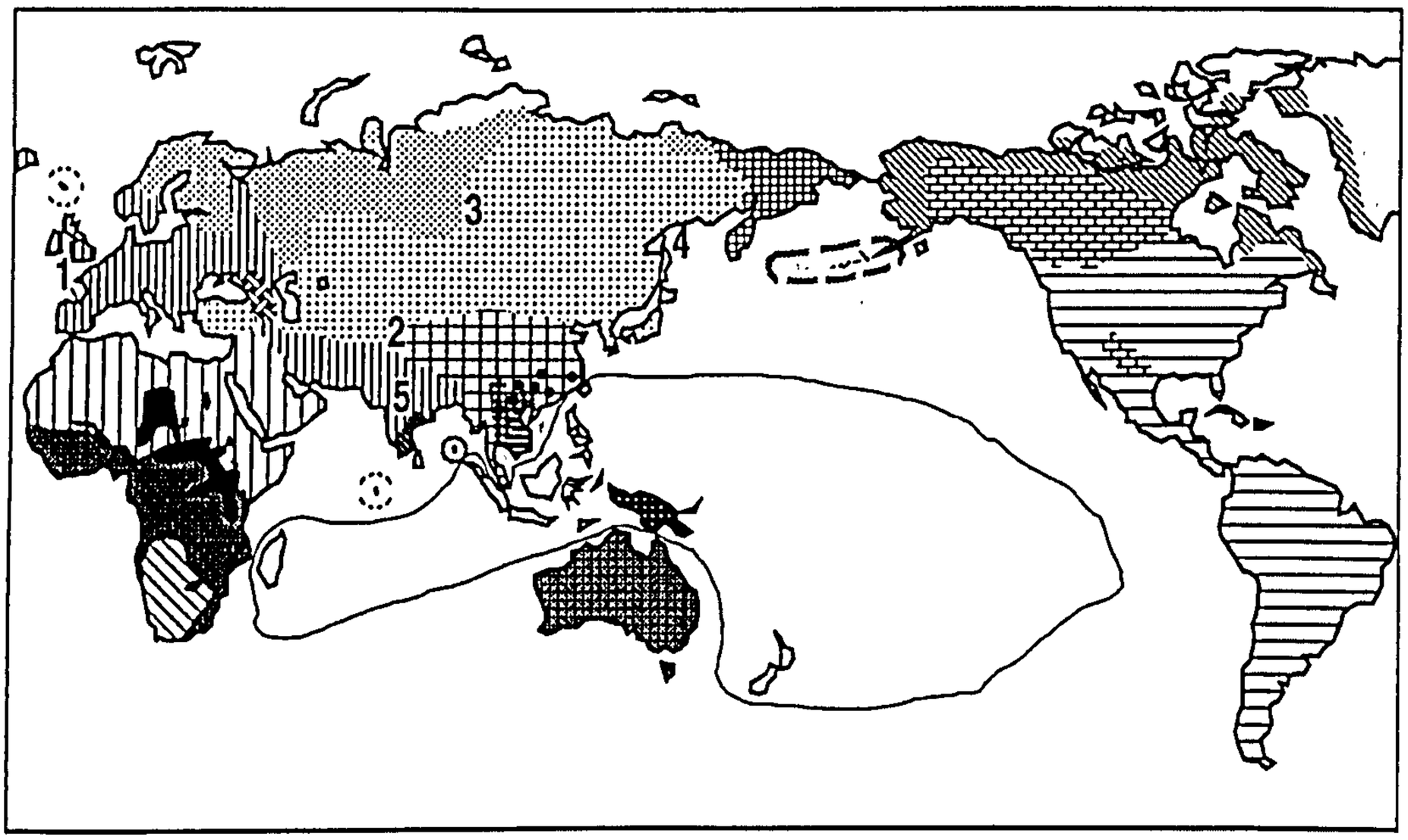

Figure 29: Merritt Ruhlen's (1987) linguistic classification.

Adapted from Cavalli-Sforza, Menozzi & Piazza, 1994:97 (Fig. 2.6.1) and Ruhlen, 1991:284-285 (Map 8.1).

Coming back to the previous citation from Cavalli-Sforza (2000), it seems that it is not the linguistic classification's acceptance by linguists which is important, but its suitability for "comparing genetic and linguistic evolutions" (p. 139):

> [...] Defining a family does not appear to be an entirely objective task, but the distinctions between families, subfamilies, and superfamilies are mostly a matter of convenience and are unnecessary for certain purposes. What matters is the possibility of establishing a *simple, logical, and hierarchical relationship.* Unfortunately, most modern classifications stop at the level of families, of which there are as many as seventeen in Ruhlen's unifying system. There are some superfamilies, but, as already noted, *modern linguistic methods have not yet generated a complete tree growing from a single source* (Cavalli-Sforza, 2000:139-140, *italics* mine).

Thus, the next logical step is made: not only using highly criticized linguistic families but trying to force them further into a "simple, logical, and hierarchical" fashion by considering

linguistic *macrofamilies*. There is a number of such proposals, but probably the best-known are *Nostratic* and *Eurasiatic*. The Eurasiatic macrofamily was proposed by Joseph Greenberg (2000, 2002) and contains Aegean/Tyrrhenian, Indo-European, Uralic-Yukaghir, Altaic, Korean-Japanese-Ainu, Gilyak, Chukotian and Eskimo-Aleut. He also tentatively connects Eurasiatic with Amerind and suggests that they represent a linguistic effect of post-LGM expansions (Greenberg, 2002). But the most frequently postulated macrofamily to date is, beyond any doubt, *Nostratic*.

A sizable body of literature is dedicated to this hypothesis, and probably its best modern appraisals are represented by the papers in Renfrew & Nettle (Eds.) (1999) and Salmons & Joseph (Eds.) (1998). The Nostratic macrofamily was first proposed by Holger Pedersen in 1903 (Bomhard, 1998:21; Pedersen, 1931) and derives from the Latin *nostrās* ["our country, native" (Renfrew, 1999:5)], but later articulated by Alan Bomhard and especially the "Moscow school": mainly Vladislav Illič-Svityč, Aharon Dolgopolsky, Vlad Dybo, Alexander Militarev and Sergei Starostin (Renfrew, 1999:4-6; Bomhard, 1998:21-23; Ramer *et al.*, 1998:61-63). There are many proposals concerning its actual composition, summarized for example in Wescott (1998) (Table 5 below).

Only Indo-European, Uralic and Altaic appear in all versions, with Kartvelian and Dravidian missing from Greenberg's Eurasiatic. It seems that one of the main divides concerns the status of Afro-Asiatic (e.g., Dolgopolsky, 1999:29): while four authors (including both the most "inclusive" - Bomhard - and the most "exclusive" - Illič-Svityč) regard it as a branch of Nostratic, two (Starostin and Greenberg) consider it a sister branch. The Nostratic tree for Bomhard (1998:27, Figure 1) is represented by Figure 30 below, while, for Starostin, following his suggestions in Starostin (1999):

> Three macrofamilies of the Old World – Hamito-Semitic [i.e., Afro-Asiatic], Nostratic and Sino-Caucasian – are quite possibly related on a deeper level. I would call the super-family uniting them all Eurasiatic (not to be confused with Greenberg's 'Eurasiatic' – which is actually a subset of Nostratic proper)
>
> (Starostin, 1999:156),

the "tree" would look like in Figure 31 below.

Author / *Group*	*Illič-Svityč*	*Dolgopolsky*	*Starostin*	*Blažek*	*Bomhard*	*Greenberg (Eurasiatic)*
Indo-European	●	●	●	●	●	●
Uralic (-Yukaghir)	●	●	●	●	●	●
Altaic (Turkic-Mongolic-Tungusic)	●	●	●	●	●	●
Kartvelian	●	●	●	●	●	
Dravidian (-Elamite)	●	●	●	●	●	
Chukchi-Kamchatkan			●	●	●	●
Japanese-Korean		●	●	●	●	●
Gilyak				●	●	●
Eskimo-Aleut			●	●	●	●
Afro-Asiatic	●	●		●	●	
Sumerian					●	
Ainu						●
Total:	**6**	**7**	**8**	**10**	**11**	**8**

Table 5: The composition of the Nostratic macrofamily as given by various authors.

Bold *= common to all 6 authors, italic = common to all except Greenberg's Eurasiatic. Adapted from Wescott (1998); further information on sources and discussion in Bengtson (1998).*

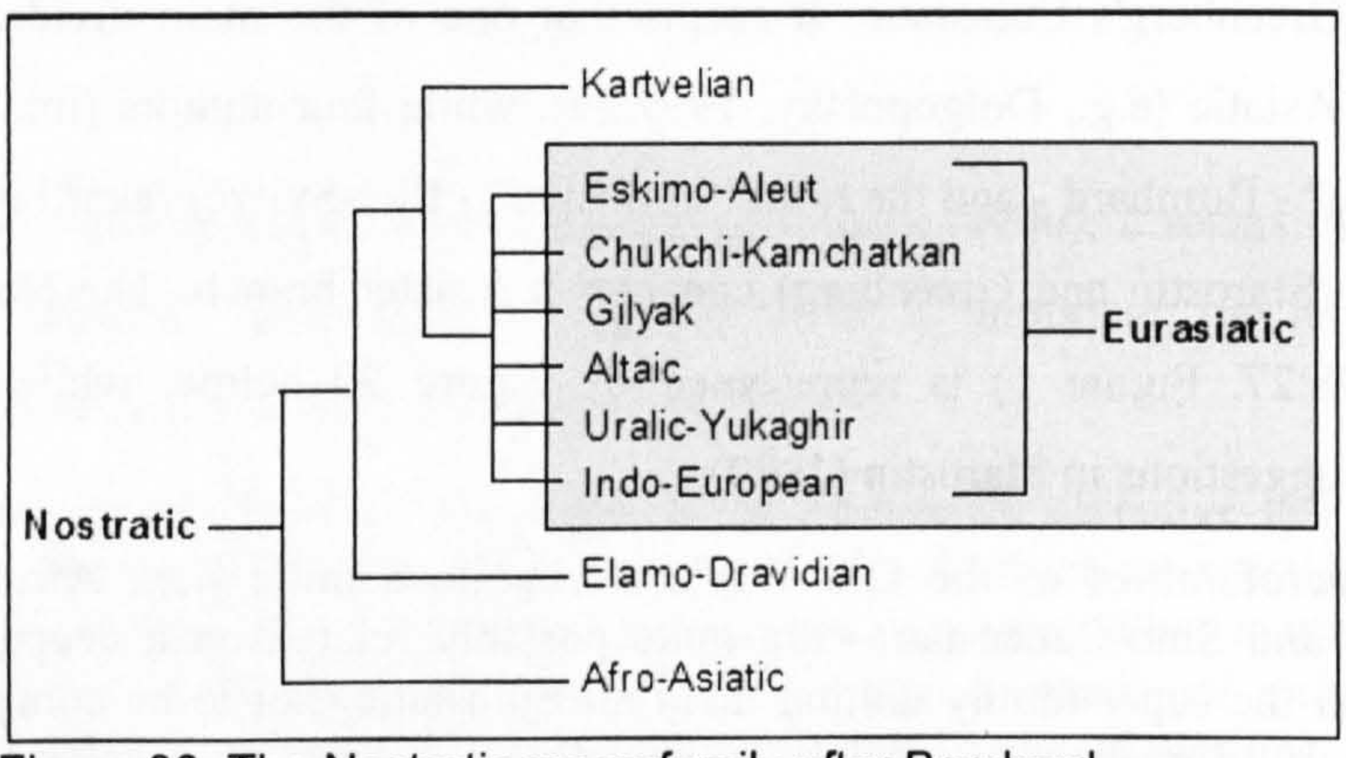

Figure 30: The Nostratic macrofamily after Bomhard.

Afro-Asiatic is a component branch, together with Greenberg's Eurasiatic (gray area). Adapted from Bomhard (1998:27, Figure 1).

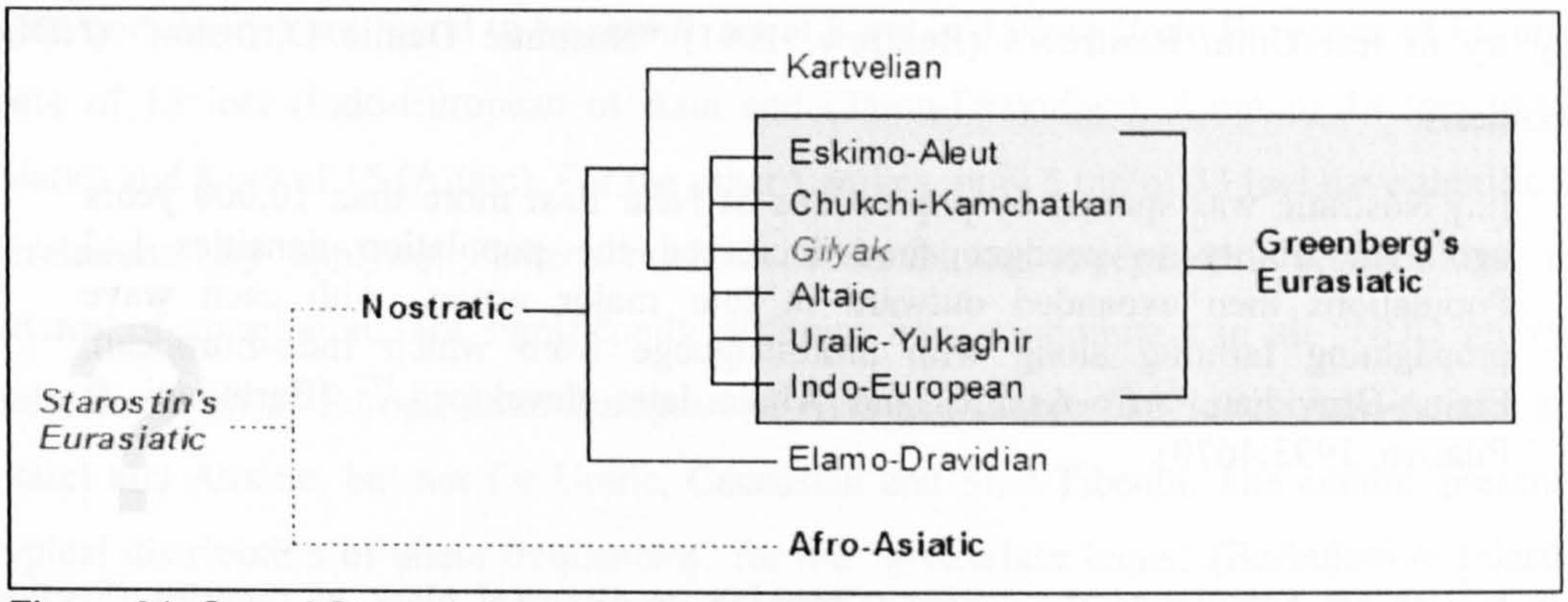

Figure 31: Sergei Starostin's version of the Nostratic macrofamily.

Afro-Asiatic is a sister branch of Nostratic inside a putative "super-family" called Eurasiatic (different from Greenberg's Eurasiatic). The status of Gilyak is not clear. Drawn using information from Starostin (1999), Renfrew (1999) and Wescott (1998).

The geographical extension of the Nostratic languages (Dolgopolsky's version) would look like:

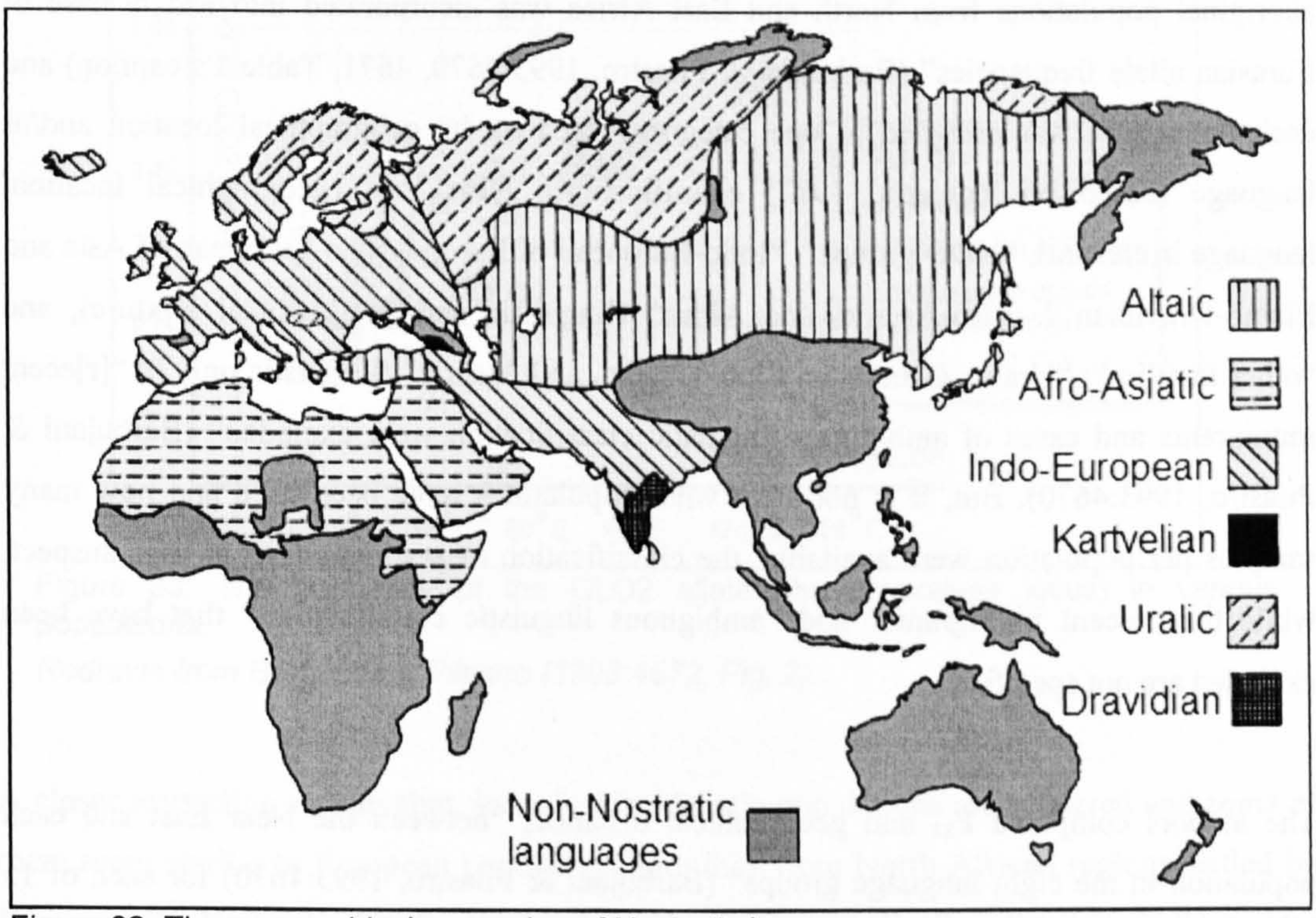

Figure 32: The geographical expansion of Nostratic languages.

Dolgopolsky's version, before the European expansions. Adapted from Renfrew (1999:6, Figure 1).

Alan Bomhard proposes "a hypothesis about possible paths by which the Nostratic sub-groups dispersed across Europe, Asia, and Africa" (Bomhard, 1998:26, see his Map 1, p. 28), which looks strikingly similar to the one proposed in Barbujani & Pilastro (1993), where

they try to test Colin Renfrew's (Renfrew, 1991) "Nostratic Demic Diffusion" (NDD) hypothesis:

> [...] Nostratic was spoken by populations of Near East more than 10,000 years ago. The ability to produce food increased the population densities [...] Populations then expanded outward in four major waves, with each wave propagating farming along with protolanguage from which Indo-European, Elamo-Dravidian, Afro-Asiatic, and Altaic later developed.[172] (Barbujani & Pilastro, 1993:4670).

This hypothesis belongs to the language/farming co-dispersal class of theories (Section 3.2.3), and it provides an extension of the simple Indo-European/farming expansion theory by adding other three "families". Their prediction is that one should find genetic gradients radiating from the Middle East not only across Europe, but in all relevant directions (Barbujani & Pilastro, 1993:4670). To this end, "information on [an unspecified set of] aboriginal populations from North and East Africa was incorporated into a data base of Eurasian allele frequencies" (Barbujani & Pilastro, 1993:4670, 4671, Table 1's caption) and each population was assigned to one category based on its geographical location and/or language (based on Ruhlen's, 1987) classification): Near East (geographical location, language irrelevant), "NDD groups": "Indo-European of Europe, Indo-European of Asia and Elamo-Dravidian[173], Afro-Asiatic, and Altaic" (linguistic and geographical mixture), and "other families": "Uralic, Caucasian, Sino-Tibetan, and Austric" (linguistic only?); "[r]ecent immigrants and cases of ambiguous linguistic classification were excluded" (Barbujani & Pilastro, 1993:4670). But, it is not clear what populations have been used and how many samples per population were available; the classification of populations is at least suspect, while the "recent immigrants" and "ambiguous linguistic classification" that have been excluded are not specified.

The authors computed F_{ST} and geographical distances "between the Near East and each population in the eight language groups" (Barbujani & Pilastro, 1993:4670) for each of 15 chosen loci and "Spearman's correlation coefficients"[174] were computed between genetic and geographic distances for each locus and group (Barbujani & Pilastro, 1993:4670-4671).

172The absence of Uralic from this version of Nostratic is to be noted (see below).

173An explanation for this strange melange of "Indo-European of Asia and Elamo-Dravidian" is offered on page 4671, Fig. 1's caption: "[...] (clumped together under the assumption that the spread of the former languages around 3000 B.C. involved negligible population replacement)".

174It is unclear why Spearman's correlation coefficient was used instead of Pearson's, but the best hypothesis is that the relationship between the two distances is not linear.

These correlations are found to be significant at 9 out of 15 loci (Indo-European of Europe), 8 out of 15 loci (Indo-European of Asia and Elamo-Dravidian), 5 out of 14 loci (Afro-Asiatic) and 8 out of 15 (Altaic). For the other families, only 5 out of 53 loci have significant correlations. By applying Fisher's method of combining independent probabilities, the patterns of correlations are significantly different from randomness in all "NDD groups" (Indo-European of Europe, Indo-European of Asia and Elamo-Dravidian, Afro-Asiatic and Altaic) and Austric, but not for Uralic, Caucasian and Sino-Tibetan. The authors present a "typical distribution of allele frequencies" for the "glyoxalase locus" (Barbujani & Pilastro, 1993:4672, Fig. 2, redrawn below in Figure 33), which they interpret as showing "approximately longitudinal clines [...] for populations speaking Indo-European, Elamo-Dravidian, and Altaic languages, but not for Afro-Asiatic speakers" (p. 4671).

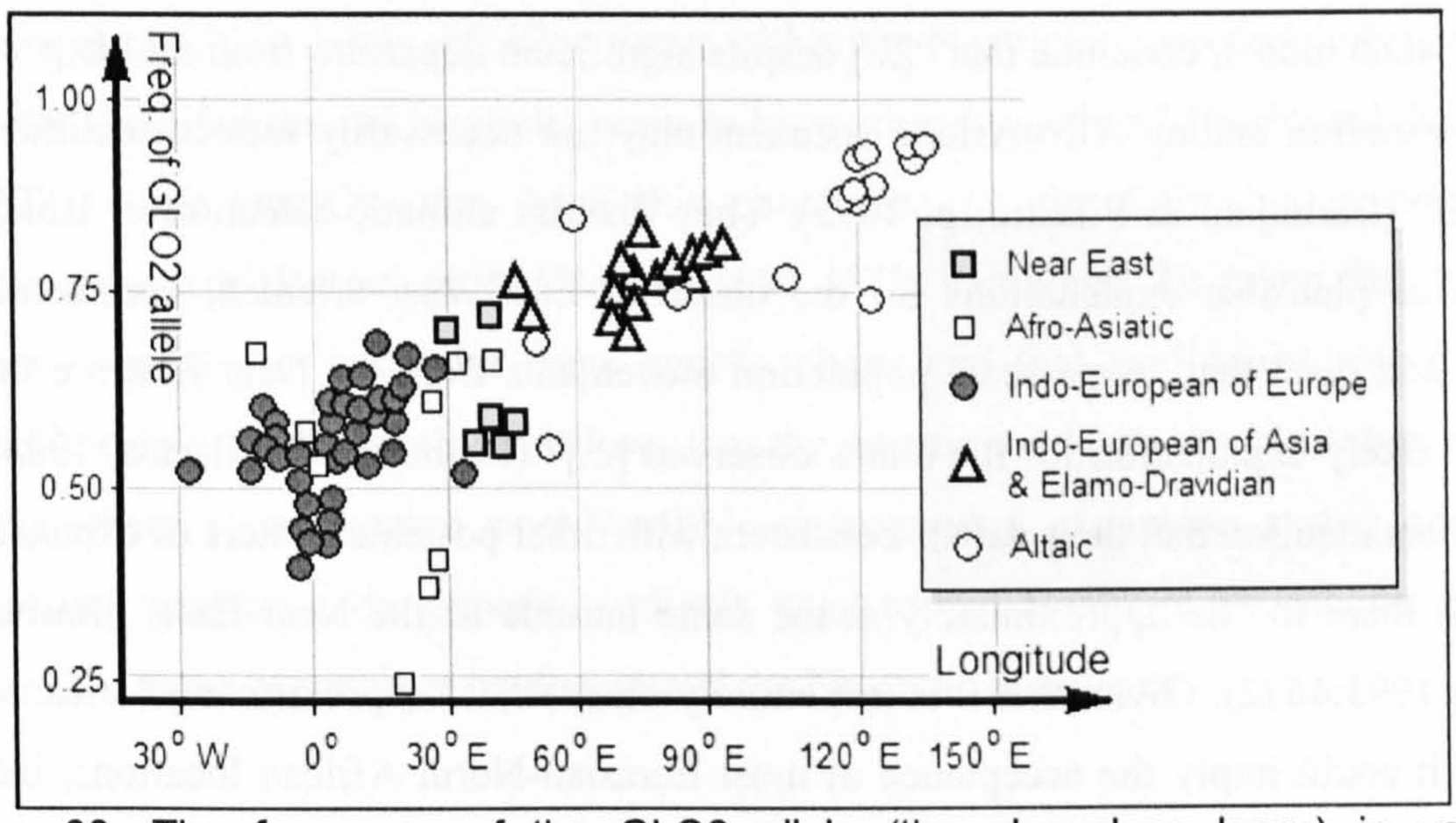

Figure 33: The frequency of the GLO2 allele (the glyoxalase locus) in various populations.

Redrawn from Barbujani & Pilastro (1993:4672, Fig. 2).

A closer inspection reveals that, indeed, Afro-Asiatic populations are scattered and some of them seem similar to European populations (are they from North African regions settled by Europeans?; Hourani, 2002), while the others seems clinally distributed from low frequencies in the West (Europe) to high frequencies in the East (Asia), with intermediate frequencies in the Near East. The most parsimonious interpretation of such a pattern is not a four-wave migration originating in the Near East, but a diffusional gradient, due to either selection or genetic drift from origin. If all loci tend to behave in this manner (East-West clines), then the four-wave migration hypothesis is superfluous and a simpler diffusional

process following the main longitudinal axis of Eurasia (Diamond, 1998:176-191), historically known to have influenced demographic processes and gene flow, would be a better explanation.

Another problem is the presence of "clines resembling those caused by the spread of alleles of Near Eastern origin" in the Austric control group (Barbujani & Pilastro, 1993:4671-4672), for which a diffusional process from the Near East could not be a valid explanation, as shown by their own dataset (Barbujani & Pilastro, 1993:4672) and other considerations concerning the origins and spread of agriculture. Therefore, they question "[...] whether some gradients in the NDD groups may also reflect processes other than neolithic demic diffusion of Nostratic speakers" (Barbujani & Pilastro, 1993:4672), and, after mentioning that there are linguists positing an African origin for Afro-Asiatic as opposed to the Near Eastern NDD model, conclude that "[...] despite significant departure from null expectations, genetic variation among Afro-Asiatic speakers may not necessarily reflect neolithic demic diffusion" (Barbujani & Pilastro, p. 4672). They dismiss climatic selection or isolation by distance as plausible explanations for the observed East-West oriented, continental-scale patterns and posit that "large-scale population movements from the Near East are therefore the most likely explanation for the clines observed [...]" (Barbujani & Pilastro, 1993:4672). The authors mention that their data is consistent with other possible centers of expansion, but constrain them to "lie approximately at the same latitude as the Near East" (Barbujani & Pilastro, 1993:4672). Given that it is not entirely clear what "approximately" means in this context, it could imply the acceptance of most Eurasian-North African locations, including for example, the Pontic steppes (Mallory, 1991), the Balkans or Egypt.

But probably the most important observation made by the authors concerns the "[...] timing of the demic diffusion process, which *cannot be inferred from allele-frequency data*." (Barbujani & Pilastro, 1993:4672, *italics* mine), which is a perfectly justified objection to applying genetic methods to linguistic and archaeological problems in general. A genetic gradient or genetic boundary, does not come with "[...] accurate time scales attached [...]" (McMahon, 2004:9): "[genetic] [c]lines do not come with dates conveniently attached" (Jobling, Hurles & Tyler-Smith, 2004:323). This criticism was most clearly articulated concerning the gradients visible in the PC1 across Europe, with one extreme in the Near East and the other in North-Western Europe and taken by Cavalli-Sforza and colleagues (Cavalli-

Sforza, Menozzi & Piazza, 1994:291, 296-301) to represent the Neolithic expansion from Anatolia (Jobling, Hurles & Tyler-Smith, 2004; Sims-Williams, 1998; Bellwood & Renfrew (Eds.), 2002). Unfortunately, all that can be said about such a genetic pattern is that it exists and that it might reflect some demographic process(es) (if enough independent loci concord): "[...] the patterns of genetic variation that we see in the world today were not caused by any single event, but instead reflect a *palimpsest*, a mosaic of events that occurred at different times and in different places" (Relethford, 2003:102, *italics* mine). The PC1 gradient across Europe could indeed represent a demic expansion from Anatolia due to agriculture, the remnant of a Palaeolithic expansion from Africa/Near East or some other, unknown event, or, most probable, a superposition of such events.

Barbujani & Pilastro (1993) conclude that, "[...] because the clines expected under the hypothesis of neolithic demic diffusion occur within the Nostratic macrofamily [...], farming and at least three families of Nostratic seem to have spread together." (Barbujani & Pilastro, 1993:4673). Unfortunately, they base this conclusion on the following contorted and fallacious argument (Barbujani & Pilastro, 1993:4673, paragraph 2): given that Nostratic languages correlate so well with these genetic clines and that no linguist would posit a Palaeolithic origin for Nostratic, it follows that the genetic gradients are Neolithic, and also given that there is no known post-Neolithic demographic expansion strong enough to generate such gradients, the genetic gradients must reflect Neolithic expansions from the Near East, the same expansions having spread the Nostratic languages.

But why did the authors use this extremely peculiar version of Nostratic, leaving aside the core components Uralic and Kartvelian (Table 5)? This is rooted in Renfrew's speculations (Renfrew, 1991), but, the exclusion of Uralic from Nostratic while keeping Indo-European is totally unjustified on linguistic grounds and is a clear example of wishful thinking. As repeatedly pointed out, of all the proposed Nostratic correspondences, Indo-European – Uralic is one of the strongest[175] (e.g., Salmons & Joseph, 1998:4; Hamp, 1998:15, Footnote 3; Greenberg, 1998:53) on linguistic grounds and its selective elimination from Nostratic is totally unacceptable. If Uralic is included in the NDD group: instead of 3 out of 4 NDD groups showing the expected gradients and 1 out of 4 non-NDD groups, we would get 3 out of 5 NDD and 1 out of 3 non-NDD. If the other unjustifiably excluded core group,

175Be it genetic or due to borrowing, see below.

Kartvelian is included in NDD (which has equal rights with Elamo-Dravidian to be considered, see Table 5), there are 3 out of 6 NDD and 1 out of 2 non-NDD groups showing the expected gradients.

So, what did in fact Barbujani & Pilastro's (1993) paper show? If we are to consider the Nostratic macro-family as it is viewed by Nostraticists themselves, we are left with a totally unconvincing 3 out of 6 (50%) NDD versus 1 out of 2 (50%) non-NDD groups showing genetic clines with one origin in the Near East. If we are to accept their peculiar and unjustified version of Nostratic, then the conclusion is that 3 out of 4 groups arbitrarily classified as "NDD" versus 1 out of 4 "non-NDD" show the genetic clines, but this simply reduces to 4 out of 8 (50%) groups showing the clines, which, again, does not support anything. Colin Renfrew tries to offer a justification for this version of Nostratic (Renfrew, 1999:10-11 and the caption of Figure 2, p. 10), by explicitly excluding Kartvelian because of "biogeographical factors [...] (i.e. the Caucasus Mountains) would prevent or limit dispersal" (Renfrew, 1999:11) and Uralic, where "[...] a later punctuation episode must be proposed [post 8000 BC – Figure 2's caption]" (Renfrew, 1999:11) by hunter-gatherers. The main problems here are that the origins of Afro-Asiatic are debated between Africa and the Near East (Renfrew, 1999:10-11; Bar-Yosef, 2002; Hassan, 2002; Militarev, 2002; Barker, 2002), Kartvelian does not show signs of expansions, and proto-Uralic speakers did not seem to have been agriculturalists – this is one of the most devastating critiques facing such studies. *Linguistic palaeontology* (Comrie, 2002:410-412; Mallory, 1991:110-127) implies that (proto-) Nostratic speakers were not familiar with agriculture (Campbell, 1999:222-223; Bomhard, 1999:68-70), which is in stark contrast with the same type of evidence, compellingly supporting agriculture for PIE (Mallory, 1991:117-120; Comrie, 2002:414-416; Fortson, 2004), Afro-Asiatic (Militarev, 2002), Dravidian (Fuller, 2002:200-205) and Austro-Asiatic (Diffloth, 2005; Higham, 2002). In this particular context, lack of evidence *is* most probably evidence of lack: (proto-) Nostratic speakers were not familiar with agriculture, implying that there could not have been the agriculturalists expanding from the Near East who carried it over Eurasia and North Africa. In conclusion, the genetic gradients detected by Barbujani & Pilastro (1993) do not provide any support for Renfrew's Nostratic demic dispersal speculation.

Moreover, from a linguistic point of view, Nostratic is an extremely problematic concept.

And given that Nostratic is *primarily* a linguistic hypothesis, no matter how much genetic and archaeological conjectural evidence is thought to support it, it is *linguists* who must be credited with its acceptance or rejection (Renfrew & Nettle, 1999; Renfrew, 1999). And, as an overview of the recent literature (Salmons & Joseph (Eds.), 1998; Renfrew & Nettle (Eds.), 1999) proves, most linguists *strongly reject* the Nostratic hypothesis, mainly on the following grounds[176]:

- **methodological**: these have the overall effect of vastly increasing the probability that correspondences ("cognates") are found due to chance only (Campbell, 1999; Ringe, 1998; McMahon & McMahon, 2005). The application of the same methodology to show correspondences between Nostratic and Nilo-Saharan and Niger-Congo (Ehret, 1999), Salishan (Shevoroshkin, 1999), Basque (Trask, 1999) and Sino-Caucasian[177] (Starostin, 1999) can be taken as the ultimate proof of the intrinsic flaws of this methodology (by *reductio ad absurdum*):
  - *non-standard attitude towards sound correspondences*: while the main difference between Greenberg and the Nostraticists is that the latter profess a strict adherence to the comparative method (Campbell, 1999; Bomhard, 1998; Greenberg, 1998), there are many critiques concerning the actual application of this method to their datasets. Their usage of sound correspondences seems to be much too lax, involving a far too liberal usage of under-specified phonetic symbols and frequent violations of the proposed sound correspondences based on special pleading (Campbell, 1999:183-188; Ringe, 1998);
  - *poor control of borrowing*: areal effects are very poorly controlled, but they could have had a very important role in shaping the current linguistic diversity (Dixon, 1997). Borrowing is a real problem for the list of "cognates" thought to support the Nostratic hypothesis (Campbell, 1999:188-197);
  - *poor control of semantic similarity*: given that Nostratic is mainly based on lexical reconstructions[178], excessive liberalism in choosing which meanings in different languages are considered similar, can profoundly increase the apparent relatedness of the families composing Nostratic. Several such "matches" are: "'root', root-crops, edible roots': 'sinew': 'stump of cabbage'; 'edible root, carrot,

176I will use Campbell (1999).
177Itself problematic.
178As opposed to, for example, Indo-European, where a very important role is played by morphological paradigms (Mallory, 1991; Fortson, 2004;)

parsnip'; 'tendon, nerve', 'tip of nose', 'muscle'; Indian horse-radish tree'; 'tendon'" (Campbell, 1999:198);

- *the inclusion of short, onomatopoeic and nursery forms*: it is known that short (monosyllabic) forms are problematic, as they can be similar simply due to chance (Ringe, 1998). Also, nursery forms are known to be similar across languages due to functional, language-external pressures (Campbell, 1999:203). Onomatopoeic forms tend also to be similar due to obvious reasons of approximating the same target sound. Campbell (1999:199, 202-203) lists a series of such forms included in the "cognate" sets proposed as supporting the Nostratic hypothesis;
- *"reaching down" and using only two component families*: for some "cognates" only two (not the same in every case) families are used, greatly increasing the probability of chance correspondences. Moreover, for some "cognates" the forms are taken from sub-families or even languages of the considered family, even if there are no agreed reconstructions of the forms to the family level ("reaching down"[179]); this is a very serious problem for Nostraticists and standard methods forbid it entirely (Campbell, 1999:200-202; Appleyard, 1999:307);
- *overlapping sets and comparison of non-cognates*: single forms in a given language are considered to belong to disjunct sets of cognates, while, by definition, a form can belong to only one set of cognates (Campbell, 1999:202). Moreover, sets of non-cognates or proposed cognates in one family are compared to sets from other family and used to support the Nostratic hypothesis (Campbell, 1999:203-204);
- *plain errors*: erroneous morphological analyses (Campbell, 1999:199) and reconstructions (Campbell, 1999:204) are also used to support the hypothesis;

- **Typology**: the proposals for proto-Nostratic have problems fitting the known typological constraints (Campbell, 1999:205; Bomhard, 1999), casting doubt about the plausibility of these reconstructions. Moreover, "[t]ypological traits which are commonplace and show up frequently in unrelated languages are not reliable evidence of genetic relationship" (Campbell, 1999:205). Nevertheless, many

179So called because if one were to represent the family tree top-down, with the proto-language on top and terminal nodes (languages) on bottom, the form is taken from the bottom even if there is no reason to assume that it actually can be reconstructed to the top.

"(macro-) families" turn out to be based exactly on such traits, e.g., Altaic or Khoisan: Altaic is included as a valid linguistic family in all Nostratic proposals to date (and in Greenberg's Eurasiatic) (Table 5), casting doubt on any "reconstructed proto-form" including "proto-Altaic";

- **Areal linguistics:** it seems highly probable that the similarities observed between languages/families included in the Nostratic hypothesis are for the most part real but due to areal effects, as argued, for example, by Dixon (1997:37-44), Campbell (1998:207-210) and Ringe (1998). In this case, the entire endeavor of trying to apply the comparative method (or a customized variant thereof) is by definition bound to fail.

It can be concluded, then, on a more pessimistic note than Daniel Nettle's closing paper (Nettle, 1999), that the Nostratic hypothesis has an extremely high probability of being simply wrong. Of course, it could well be that linguistic families proposed as components of this macro-family are indeed very remotely related, but I tend to agree with Dixon (1997) that at such time depths areal effects might become increasingly important, voiding the question of genetic relatedness of any meaning. If Nostratic, by far one of the best studied and methodologically sound of the proposed "macro-families", has such dim prospects, one can conclude that, at least for the moment, it is better to avoid considering such linguistic constructs in any respectable interdisciplinary work.

3.2.4.3. The concept of "population" and sampling problems

Any research involving the study of genetic diversity, involves a sampling strategy concerning human groups. Probably the best sampling strategy imaginable would be:

> Ideally, a random sampling procedure based on a physical grid approach should be employed, but such intensive structured sampling has not been performed in human research to date, and is unlikely to form a significant part of future research for socio-political as well as scientific reasons (McMahon, 2004:4).

Such a sampling, based on objective, predefined criteria, would allow the collection of genetic frequency data for human groups living on a regular grid, equally spaced by the same distance, and irrelevant to external criteria like political affiliation, ethnic labels, economic affluence, population size or subjective valuing of different "ethnic groups". Unfortunately, such a research program is doomed to failure due to the enormous costs and logistic

difficulties involved. Thus, in general,

> [...] sampling has often been based on the basis of named, culturally significant groups, such as villages or ethnic groupings defined by language affiliation, and small, disappearing tribal groups characterized on the basis of their language are often treated as equivalent to similar-sized samples drawn from large, modern nation states (McMahon, 2004:4).

Cavalli-Sforza, Menozzi & Piazza (1994:20-22) try to justify such an approach and observe that "[...] in practice, one deals with samples that have already been collected and tested so that one is limited to deciding whether a sample is acceptable" (Cavalli-Sforza, Menozzi & Piazza, 1994: 20), which is extremely important and valid. But besides this objective limitation, their methodology does suffer from a series of major flaws (McMahon, 2004:4-6; Bateman *et al.,* 1990:2-4; Sims-Williams, 1998; esp. MacEachern, 2000:360-363).

The samples used are not equivalent in the sense that

> [t]he groups that form the basis of these analyses [i.e., this sampling strategy] can thus range from very small, marginalized communities only weakly integrated into modern political and economic systems through extremely large and complicated ethnic units to the citizenry of national states (MacEachern, 2000:361),

like the Hadza of Tanzania (population of ~1000), South Chinese (population ~500 million) and French (population ~60 million), which are considered as equivalent samples (MacEachern, 2000:361). This worry is entirely justified, given that genetic drift is highly dependent on population size (Halliburton, 2004:221-265; Jobling, Hurles & Tyler-Smith, 2004:131-137) and there is no a priori reason to expect that the degrees of admixture and diversity are the same across such a range of population sizes and socio-political organizations. For example, the histories of Europe (Davies, 1997) or the Near East (Hourani, 2002) show beyond doubt the amount of population movement and admixture witnessed by the last thousands of years in these particular cases, facilitated by specific cultural, economic and political factors. Thus, a "French" sample is potentially not equivalent on genetic grounds to a "Hadza" sample[180].

There is also the assumption that

> [e]xcept for the very few widely spoken languages, there tends to be a one-to-one correspondence of tribal names to language names. Thus, except in the case of

180This is not to deny that such small populations (like the Hadza) could also experience tremendous amounts of admixture. For a very interesting example of ancient Jewish male admixture into a South African tribe (Lemba) see Bradman, Thomas, Weale & Goldstein (2004).

> large modern nations in which the identity of original tribes is usually – though not entirely – lost, *languages offer a powerful ethnic guidebook*, which is essentially complete, unlike strictly ethnographic information (Cavalli-Sforza, Menozzi & Piazza, 1994:23, *italics* mine),

but, this assumption is not only simplistic, it is totally misleading (Sims-Williams, 1998:517-519). For example, detailed studies of the Yanomama tribes

> [...] have shown a high degree of fission and fusion, intermarriage and warfare amongst the roughly 150 villages that make up this linguistic group, and recent results seem to indicate that *several villages are genetically closer to geographically close, but linguistically and culturally distinct groups*, than they are to other Yanomama villages, either due to higher rates of gene flow or shared common genetic but not cultural ancestry (McMahon, 2004:5, *italics* mine).

Moreover, bi- and multi-lingualism represent the norm and not the exception (Dixon, 1997; Sims-Williams, 1998:517), which

> [...] is not a trivial criticism, because *bilingualism is the prerequisite for language-shift*, a phenomenon which has occurred on a massive scale not only in modern times [...] Cavalli-Sforza and colleagues underplay the importance of language-shifts, and of language deaths [...] (Sims-Williams, 1998:517, *italics* mine),

and there are cases where "[...] language differences determined by linguists may not even match the boundaries deemed to be important by the tribal groups themselves" (McMahon, 2004:5). There is a generally held misconception about language shift, namely that it is infrequent, that it is a somehow "unnatural" phenomenon (Dixon, 1997; Trask, 1999; Sims-Williams, 1998), exemplified by Cavalli-Sforza and colleagues: "Language replacement is more likely to happen, perhaps, *in recent history*, and there are well-known examples of it" (Cavalli-Sforza, Menozzi & Piazza, 1994:157, *italics* mine).

A striking example of the distorting effects these assumptions can have is illustrated by Africa:

> For Africanists, however, even more striking may be the uncritical acceptance of dated Western models of "tribal" or "ethnic" identification [...] It is abundantly clear that many of the "tribes" so beloved of (even modern) Western commentators are not entities preserved unchanged from ancient times but rather the relatively recent products of intense participation in regional networks of political, social and economic interaction (MacEachern, 2000:362),

and he shows how these "tribes" are artificial creations due to the need of easy administration and government (MacEachern, 2000:363).

The sampling procedure used by Cavalli-Sforza and colleagues is thus open to debate and

these critiques can be applied to most such studies. Unfortunately, given that there already exists a sizable database of such samples and the high costs involved in a new, more principled, sampling program, one has to address these problems and try to minimize their impact on each study individually. Therefore, the potential biases induced by such sampling strategies must be remembered at all times, especially when parallels between population and linguistic "classifications" or "distances" are claimed.

3.2.4.4. Parallels between linguistic and genetic classifications

Probably the most criticised aspect of Cavalli-Sforza and colleagues work concerns the comparison between linguistic and genetic classifications. They used a tree classification of human populations based on genetic distances, resulting in a phenetic populations tree (see above), which was compared to their preferred linguistic classification, concluding that "[t]he one-to-one correspondence between genetic clusters and linguistic families is *remarkably high*, but is not perfect" (Cavalli-Sforza, Menozzi & Piazza, 1994:99). The comparison between the two classifications is represented in Figure 34. The gray boxes represent linguistic entities not generally accepted by linguists: Amerind (with its subdivisions, South, Central and North; Greenberg, 1987), Indo-Pacific (Greenberg, 1971), Australian, Austric (Reid, 2005), Altaic and Eurasiatic, and Nostratic.

Concerning the "one-to-one correspondence between genetic clusters and linguistic families" (Cavalli-Sforza, Menozzi & Piazza, 1994:99), Figure 34 seems at first sight to show a striking parallelism, but a closer inspection reveals, first, that the linguistic classification is *not* hierarchical, but simply a list of linguistic families (Bateman *et al.*, 1990:6). Therefore, Figure 34 becomes Figure 35 where the correspondences are greatly diminished. But even the remaining ones (e.g., Niger-Kordofanian, Indo-European, or Austronesian) are illusory, based on the exploitation of "[...] the mobile-like properties of a repeatedly branching tree [...]; nodes of the phenogram are rotated to achieve maximum *apparent* congruence of populations and linguistic phyla" (Bateman *et al.*, 1990:6, *italics* in original).

This is exemplified by a quick analysis of the Indo-European family: from the depiction, it looks like "Iranian", "European", "Sardinian" and "Indian" populations form a cluster of Indo-European languages, to the exclusion of the others. But the inspection of the tree

reveals that "Iranian" is phenetically closer to "S.W. Asian", which, together with "Berber, N. African" and "Ethiopian" speak Afro-Asiatic languages. Moreover, "European" is equally close phenetically to the cluster "Iranian" + "S.W. Asian" and "Sardinian" is quite remote from "Indian", which has as its phenetically closest neighbor "S.E. Indian" speaking Dravidian languages. A perfectly equivalent depiction of the same phenogram using a different (and more systematic) ordering of the branchings is in Figure 36.

The apparent parallelism has gone. The

> [d]etermination of *true* congruence is hampered by the non-hierarchical treatment of the linguistic phyla. In these circumstances, congruence is most appropriately assessed by observing whether a particular linguistic phylum corresponds with inclusive clusters of populations on the phenogram (Bateman *et al.*, 1990:6, *italics* in original),

which is the tendency of populations close together in the phenogram to belong to the same linguistic family. Their evaluations are 56% correspondence at a coarse level (six population aggregates) and only 11% correspondence at a more rigorous level (populations) (Bateman *et al.*, 1990:6). Moreover, "[n]either of the two linguistic superphyla (Nostratic and Eurasiatic) precisely corresponds with any of the population aggregates or groups of population aggregates" (Bateman *et al.*, 1990:6).

Another potential problem is that language is already essential in delimiting the sampled populations, following from the "language as an ethnic guidebook" principle. As McMahon (2004) warns, "[...] when we are asking questions about the relationships between human groups and their languages, to base the sampling criteria in one domain on data from the other automatically weakens the importance of any relationships detected" (p. 4).

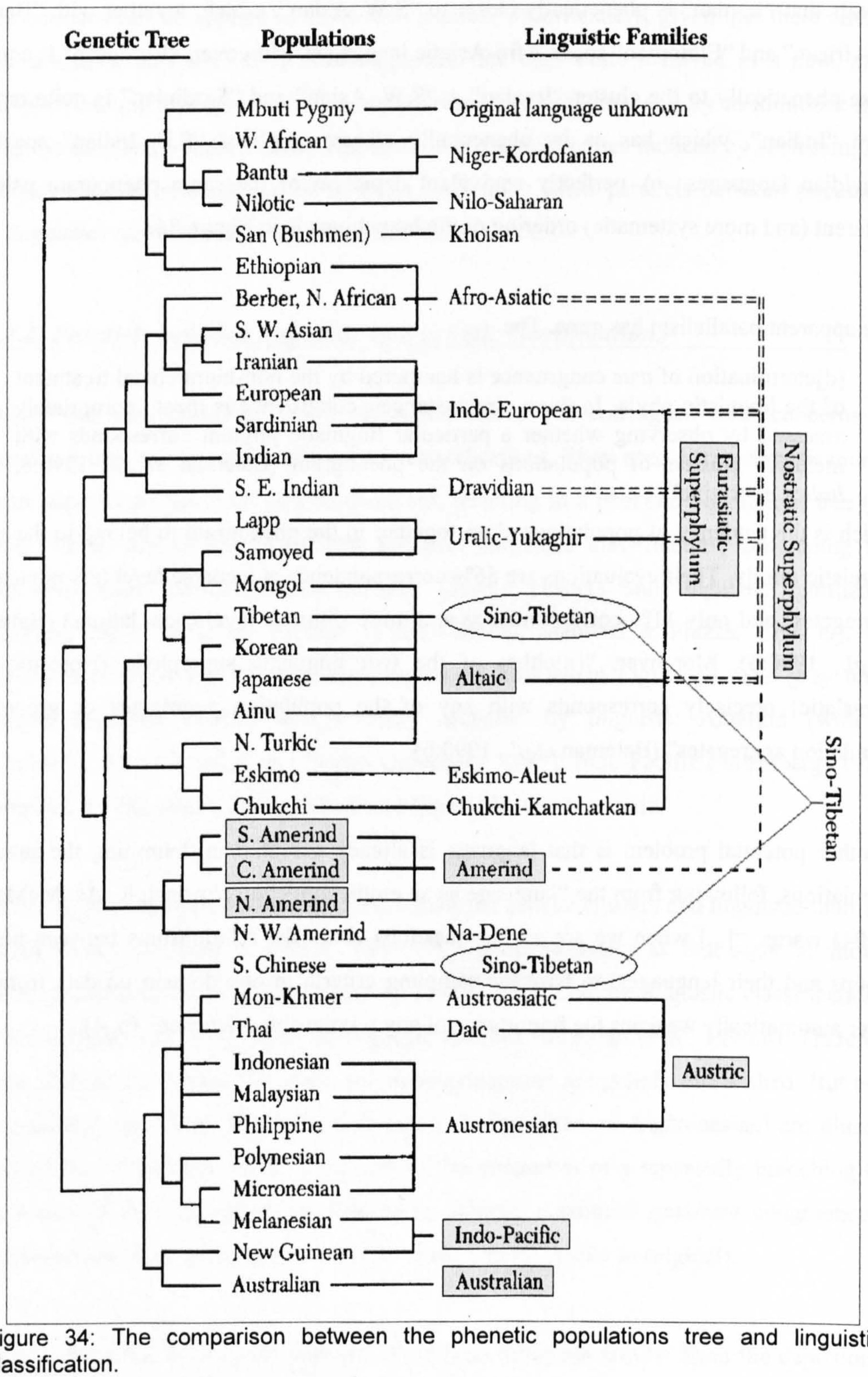

Figure 34: The comparison between the phenetic populations tree and linguistic classification.

Adapted from Cavalli-Sforza (2000:144, Figure 12) and Cavalli-Sforza, Menozzi & Piazza (1994:99, Fig. 2.6.2). Gray boxes: linguistic entities not generally accepted by linguists. Light gray ellipse: the Sino-Tibetan "split". See text for details.

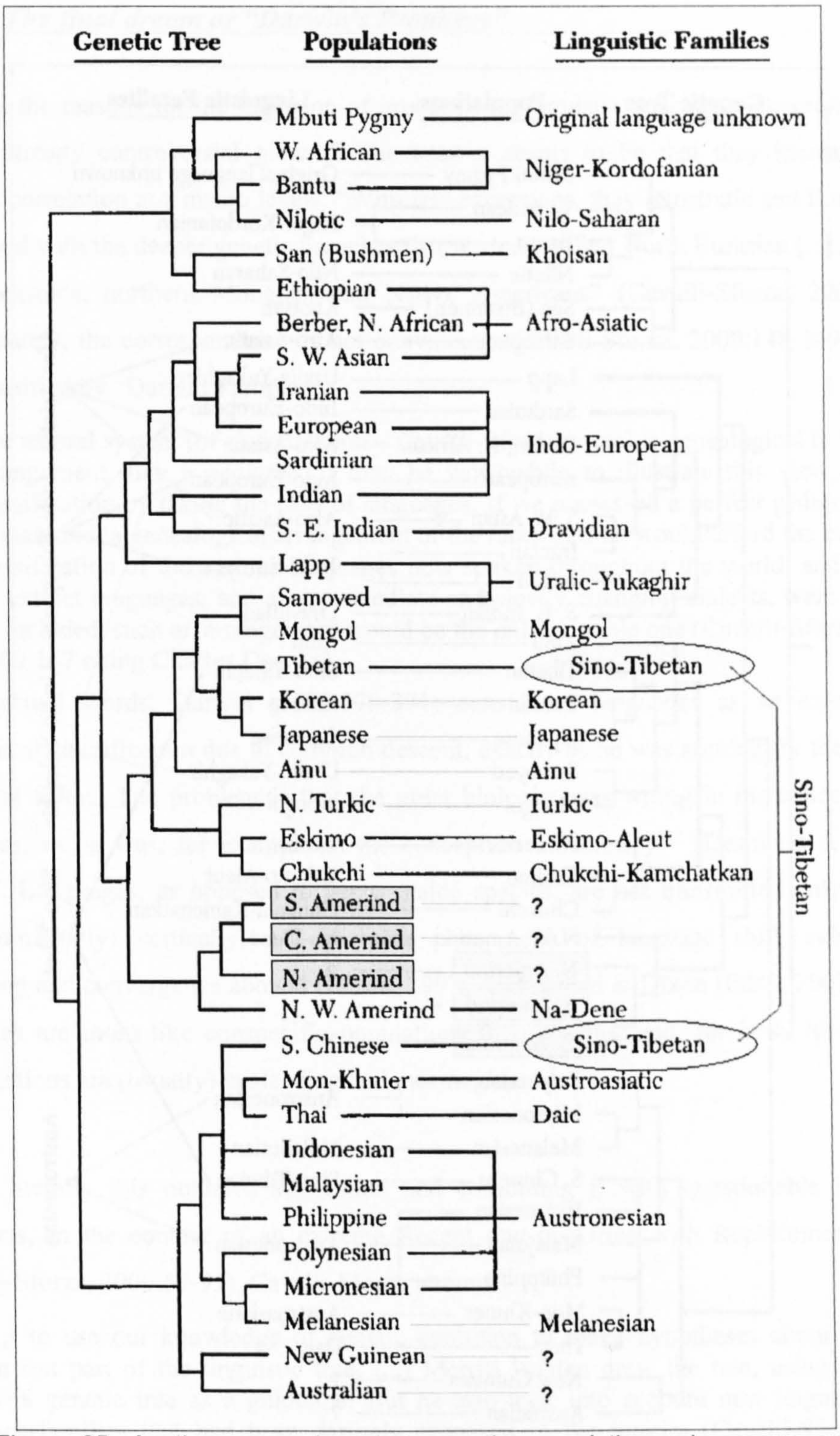

Figure 35: Another comparison between the populations phenogram and linguistic classification.

The same as Figure 34 but after the deletion of linguistically contentious entities. See text for details.

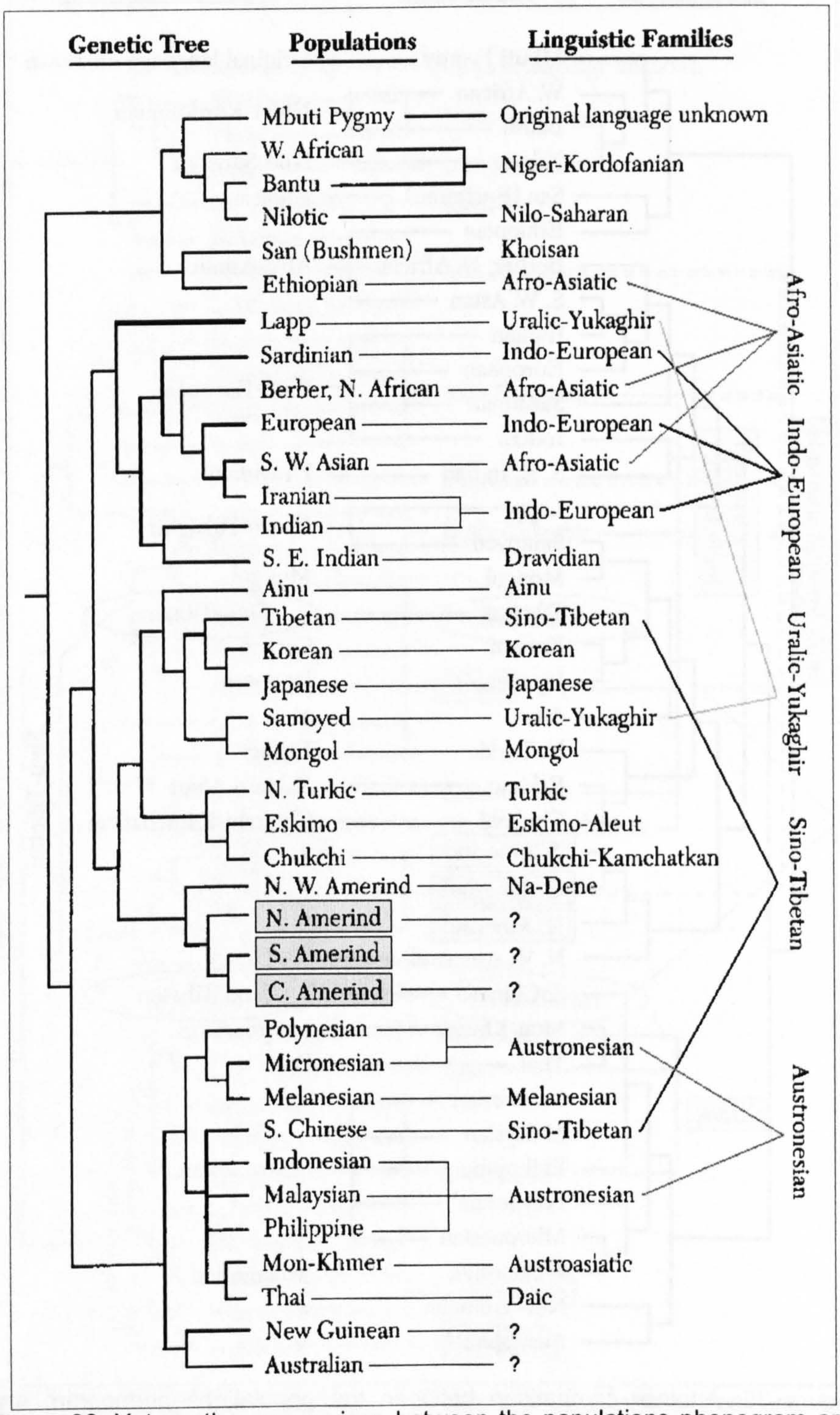

Figure 36: Yet another comparison between the populations phenogram and linguistic classification.

Same information as in Figure 35 but a different layout of the populations phenogram. See text for details.

3.2.4.5. The final dream or "Darwin's Prophecy"

What are the reasons for the inclusion of problematic linguistic constructs (macrofamilies) into an already controversial picture? The answer seems to be that they increased the apparent correlation at a macro levels: "With few exceptions, they [Nostratic and Eurasiatic] correspond with the deeper genetic branches that we have called North Eurasian [...], uniting the Caucasoids, northern Mongols, and Native Americans" (Cavalli-Sforza, 2000:147). Unfortunately, the correspondence is not convincing (Cavalli-Sforza, 2000:148-149), but it seems justified by "Darwin's prophecy":

> The natural system [of classification – Cavalli-Sforza's note] is genealogical in its arrangement, like a pedigree. It may be worthwhile to illustrate this view of classification by taking the case of languages. If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one (Cavalli-Sforza, 2000:167 citing Charles Darwin).

In his actual words, Darwin (1872:370-371) considered languages as an example of hierarchical classification due to common descent, exactly as he was arguing for the case of biological forms. The problem is that the great biologist was wrong in his conception of languages, as he was, for example in his conception of heredity[181] (Desmond & Moore, 1992)[182]. Languages, as opposed to genes inside species, are not transmitted only (or not even dominantly) vertically, and complex phenomena of language shift, substratum, borrowing and convergence abound (Dixon, 1997; Aikhenvald & Dixon (Eds.), 2001). Thus, languages are more like conspecific populations than species, and, for both, hierarchical classifications are (usually) misleading and meaningless.

Taking literally this outdated suggestion and combining it with questionable linguistic constructs, in the context of an extreme Recent Out-of-Africa with Replacement model (Cavalli-Sforza, 2000:57-91), Cavalli-Sforza attempts

> [...] to use our knowledge of genetic evolution to make hypotheses about the earliest part of the linguistic tree. [...] Merritt Ruhlen drew the tree, using our 1988 genetic tree as a guideline. But he also took into account new linguistic superfamilies that had been daringly proposed in the interim (Cavalli-Sforza, 2000:168),

---

181A mixture of blended inheritance and transmission of acquired characteristics.

182This is not, of course, intended to deny in any way the profound impact of his genial work on the modern world.

and the result is reproduced in Figure 37:

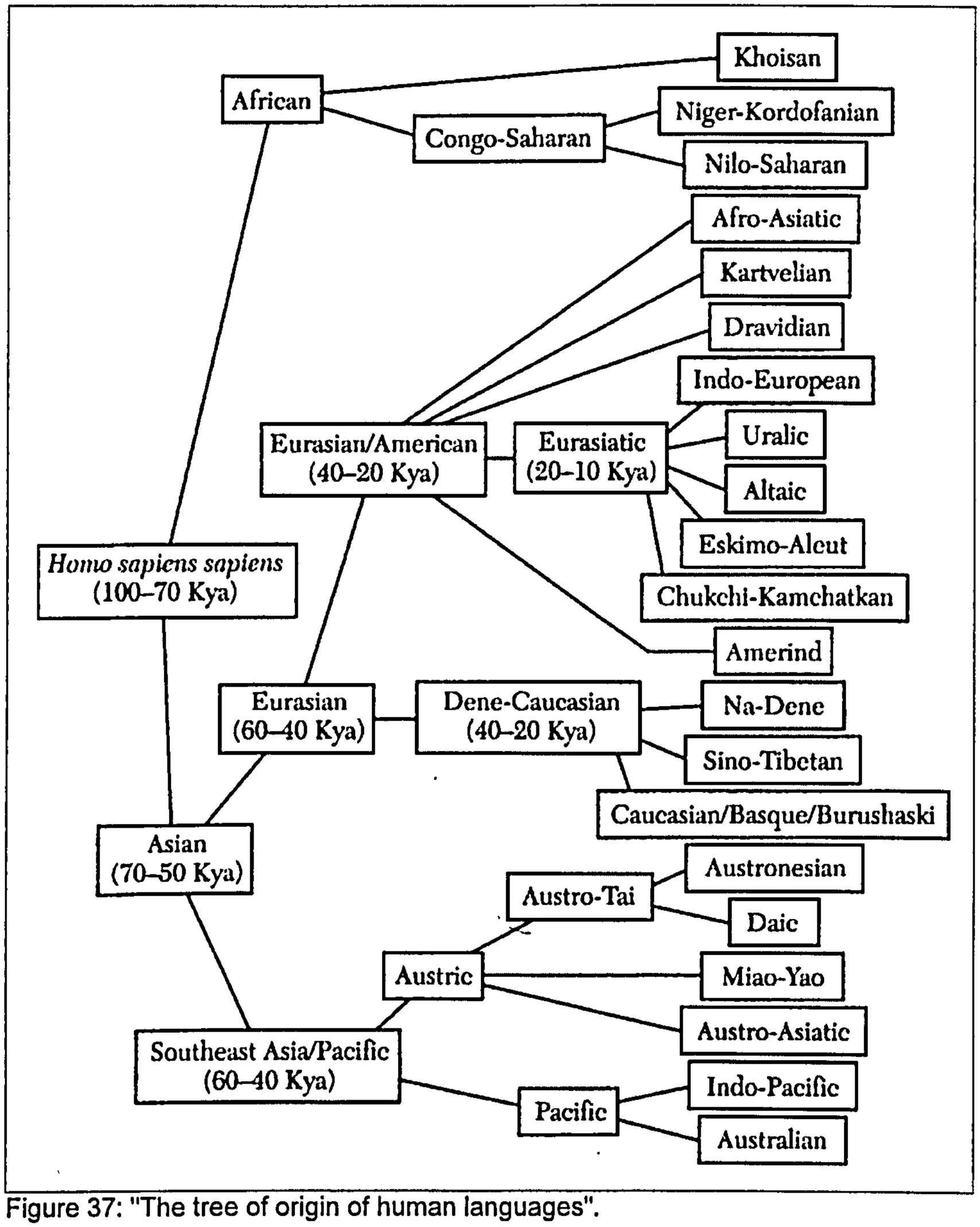

Figure 37: "The tree of origin of human languages".

Drawn by Merritt Ruhlen and modified by Cavalli-Sforza. Divergence dates are included in the parentheses. Reproduced from Cavalli-Sforza, 2000:169, Figure 14.

As bold as this might be, it remains pure speculation, and a very improbable one.

3.2.4.6. Comparing genetic and linguistic distances

Another approach to comparing linguistic and genetic patterns is represented by the computation of the correlation between genetic and linguistic distances or similarities. This generic methodology was applied, for example, by Poloni *et al.* (1997) to a world-wide[183] Y-chromosome dataset, by Rosser *et al.* (2000) to an European Y-chromosome dataset, by Sokal, Oden & Thomson (1992) to the Indo-European origins problem and by Nettle & Harriss (2003) to a West African and Eurasian database of classical markers. In principle, it involves the computation of the genetic distances between populations and linguistic distances between the languages allegedly spoken by them, so that the resulting matrices can be analyzed for common patterns (e.g., boundaries) or correlations (Mantel test). While the computation of genetic distances is a well analyzed domain, with its standardized methods and known problems (Jobling, Hurles & Tyler-Smith, 2004:166-170, 185-194; Cavalli-Sforza, Menozzi & Piazza, 1994:29-30, 39-52; Halliburton, 2004), the linguistic counterpart is fraught with important difficulties.

The first approach to obtaining such linguistic distances can be called "subjective judgment" and is exemplified by Sokal, Oden & Thomson (1992), where "[l]inguistic distances (LAN) were subjective estimates furnished by M. Ruhlen, based on his current classification of IE languages [Ruhlen, 1987]" (Sokal, Oden & Thomson, 1992:7669). This approach offers some advantages (transferring the decisions to an "expert" and fine-grading of the distances[184], allowing their treatment as interval statistical variables), but these are far outweighed by the disadvantages. The distance scheme proposed by any single linguist cannot claim to be objective in any form, and Merritt Ruhlen's scheme in particular is expected to be especially controversial. A possible quick fix would have been to obtain such judgments from a pool of linguists, including exuberant macrofamilies fans like J. Greenberg and M. Ruhlen and more orthodox ones, like L. Campbell and L. Trask, and combine them, if this proves possible. A close inspection of the tree of IE languages generated from this linguistic distances matrix (Sokal, Oden & Thomson, 1992:7671, Fig. 1) is generally in agreement with accepted classifications of IE (Fortson, 2004:8-12, esp. Figure 1.1, p.10; Mallory, 1991:9-23, esp. Figure 5, p. 15), but there are points of disagreement. For example,

183Heavily skewed towards Africa, Europe and South-West Asia, with very poor coverage of the rest of the world (Poloni *et al.*, 1997:1017, Figure 1).

184From their Fig. 1 (Sokal, Oden & Thomson, 1992:7671) it can be deduced that the distances ranged from ~1.0 to ~30.0.

in their tree, the cluster ((Danish, Swedish, Norwegian), Icelandic, Faroese) seems at odds with the historic and linguistic facts (Fortson, 2004:300-309, 328-332), which suggest the clustering ((Swedish, Danish), (Norwegian, (Faroese, Icelandic))). Thus, this method cannot be considered objective nor easily extended to other sets of languages[185].

Another method is to derive a measure of linguistic distance from a linguistic tree. Given a linguistic tree and two languages, their distance is considered to be a measure of "how close" they are in the tree. It is used, for example, by Nettle & Harriss (2003):

> The relationship between the languages of all pairs of populations were classified according to the following numerical scheme: 1, same language; 2, languages in the same branch of a family; 3, languages in different branches of same family; or 4, languages not demonstrably related. Only family relationships accepted by the consensus of historical linguists were admitted (Nettle & Harriss, 2003:334),

and by Poloni *et al.* (1997):

> [...] they [the dissimilarity indices between languages] were computed as follows: two populations within the same language family are set to a distance of 3 if they belong to different subfamilies; their distance is decreased by 1 for each shared level of classification – up to three shared levels, where their distance is set to 0. The linguistic distance was not refined any further at the interfamily level [...]. Finally, because the evolutionary distances between language families are still largely unknown but assumed to be important, a dissimilarity index of 8 was arbitrarily assigned to any pair of populations belonging to different language families (Poloni *et al.*, 1997:1017-1018).

The method can be justified on linguistic grounds, in the sense that languages belonging to the same classificatory level are descended from a more recent common ancestor than languages sharing a more inclusive classificatory level, but any tree representation is necessarily simplifying. This scheme gives the same weight to equivalent levels, but this depends crucially on the particular classification used. For example, the distance between Norwegian and Albanian, and Romanian and Albanian would be the same, but there are opinions positing a Thracian substratum in Romanian, shared with Albanian (Ciorănescu, 2002; Ivănescu, 2000), which would modify the distances. Better known is the case of English, whose distance to Dutch would be much less than its distance to French, while in reality its situation vis-a-vis French is much more complex (Ostler, 2005:456-477). Another problem is represented by the high granularity of the distance measure: in the the case of Nettle & Harriss (2003) it has just 4 levels, while for Poloni *et al.* (1997) there are 5 levels,

185 I find it hard to imagine this method applied to a set including, for example, West African, Papua-New Guinean highlands and Central American languages.

in both the last one being a bin containing all unrelated languages.

The third approach can be named "numerical" and involves the computation of distances between pairs of languages based on some characteristic(s) of the languages in question. For example:

> Within IE languages, linguistic distances were adapted from Dyen *et al.* (1992), who used the lexicostatistical method of Swadesh (1952) on comparisons of 200-word lists: percentage similarities were first converted to dissimilarities, and these numbers then assigned as nonpercentage distances between languages [...] (Rosser *et al.*, 2000:1531).[186]

The lexicostatistical method is based on an idea first introduced by Morris Swadesh (Swadesh, 1952), whereby a standard list of words chosen so that they are very resistant to borrowing or innovation and forming the so-called "core" or "basic" vocabulary is collected in as many languages as possible and judgment about their cognate status is produced by standard historical linguistic methods. There are several versions of such lists but probably the most popular is provided by the 200 words list (Dyen *et al.*, 1992). It is argued that the percent of shared cognates out of these 200 words between two languages reflects their degree of historical relatedness. For example, using the Dyen *et al.* (1992)'s dataset[187], the cognates percentage between Icelandic and Faroese is 92.2%, between Icelandic and Danish is 77.9%, while between Icelandic and Albanian is only 10.8%, suggesting that Icelandic is very closely related to Faroese, close to Danish but remotely to Albanian. From such cognates percentages data, considered as similarity measures, linguistic classifications ("phenograms") can be built (Dyen *et al.*, 1992; McMahon & McMahon, 2005). These distances are established by applying standard historical linguistic methods for judging the cognation of any two corresponding words in any two languages and involves an enormous amount of work. Also, lexicostatistics is different from glottochronology (Dyen *et al.*, 1992; Embleton, 2000; Lohr, 2000; Blust, 2000; Matisoff, 2000), which takes the cognation percentages provided by it and, assuming a radioactive decay-like rate of core vocabulary replacement model and some calibration points[188], attempts to put absolute dates on language splits. While lexicostatistics is usually considered an acceptable approximation in some cases, like Indo-European, but not applicable to others, like South-East Asia,

---

186Their full methodology is a mixture between this and a variant of the "subjective" method for Altaic and Uralic languages (Rosser *et al.*, 2000:1531-1532)

187Online http://www.ntu.edu.au/education/langs/ielex/HEADPAGE.html, September, 2006.

188Like the known divergence date for Romance languages from Vulgar Latin, begun ~2kya.

glottochronology is rejected by most linguists, despite some attempts at rectifying its major flaws (e.g., Starostin's (2000) "root glottochronology").

Other methods for computing linguistic distances between languages based on the same core vocabulary as lexicostatistics but using different principles exist. The main idea is that the more similar the lists are, the more related the languages should be. This similarity can be computed using information theoretical (Shannon, 1948) approaches: one idea is to compare word lists as a whole using the general purpose compression algorithm *zip* (Ziv & Lempel, 1977; Benedetto *et al.*, 2002). Another idea is to compute, for each list, the distances between all the component words[189], obtaining a matrix of distances, from which a confusion probability between any two words in the same list is derived. The matrices of confusion probabilities for any two languages are compared using Fisher divergence, resulting in the actual distances between languages (Ellison & Kirby, 2006)[190]. One could compare texts written in different languages, for example, translations of the Universal Declaration of Human Rights (currently translated in 365 languages[191]), or the Bible, using zip distance (Benedetto *et al.*, 2002) or other methods.

This type of approaches to computing linguistic distances has a series of clear advantages, including objectivity, repeatability and the fact that the distances produced are (from a statistical point of view) interval variables, allowing thus the usage of very powerful and complex statistical techniques. The major drawback is that they usually rely on a set of simplifying assumptions, which are not met by the messy processes of linguistic evolution, and the effects of their violations on the reliability of the computed distances is mostly currently unknown (McMahon & McMahon, 2005). Another major problem is the reliability of the data, especially for less well-known languages.

After obtaining the genetic and linguistic distances between populations, various statistical procedures can be used to assess the relationships. The most used method is the Mantel (partial) correlations test (Mantel, 1967), which computes a correlation coefficient between

189Using, for example, the string-edit or Levenstein distance (Levenshtein, 1965), but other string distances can be used [e.g. longest common subsequence, n-grams distance, dice, etc. (Kolatch *et al.*, 2004; Kondrak, 2002)].

190My own research (unpublished) seems to show that Fisher divergence coupled with the Levenstein string-edit distance fares best for IE.

191http://www.unhchr.ch/udhr/index.htm; http://www.unhchr.ch/udhr/navigate/alpha.htm

two matrices or a partial correlation coefficient between two matrices when controlling for the effects of a third (Fortin & Dale, 2005:147-153; Bonnet & Van de Peer, 2002). Sokal *et al.* (1992) computed the Mantel partial correlation between genetic and linguistic distance matrices when controlling for geographical distances and obtained a small number of low positive correlations (average across loci of 0.059, only 7 out of 25 significant) (Sokal *et al.*, 1992:7670), confirming that the main cause of language-genes correlation is geography, with a very low residual correlation between them (Sokal *et al.*, 1992:7671). Poloni *et al.* (1997:1018) also computed Mantel (partial) correlations involving genetic, linguistic and geographic distances and obtained that genetic and linguistic distances correlate strongly and significantly ($r = .588$, $p < .001$), while the amount of genetic variance explained by geography (37.6%), linguistics (32.1%) and both (44.1%) supports the hypothesis that language and especially geography strongly influence genetics (Poloni *et al.*, 1997:1021 and Table 2, p. 1020). This study refers exclusively to the Y chromosome and the .567 highly significant correlation found between genetic and linguistic distances is partially accounted for by geography. Another application of Mantel (partial) testing to linguistic, genetic and geographical distances for the Y-chromosome in Europe is offered by Rosser *et al.* (2000), which obtained that the partial correlation between genetics and language when controlling for geography is not significant, while that between genetics and geography when controlling for language was significant ($r = .349$, $p < 0.001$), confirming that the main explanation for the pattern of Y chromosome genetic diversity in Europe is offered by geography (Rosser *et al.*, 2000:1537, 1540).

A related but somehow convoluted method was used by Nettle & Harriss (2003), which, after computing the simple correlation coefficients between pairs of distances[192], have performed linear regression of (logged) genetic distance on geographic distance, sorted the residual genetic distances by the degree of linguistic relationship (4 classes in their case) and performed ANOVA for each region. They found that there is a general trend within regions for the residual genetic distance to decrease with increasing genetic relatedness (especially clear in Europe) (Nettle & Harriss, 2003:334-335, Figure 1, p. 336). In other words, what Nettle & Harriss (2003) do is try to relate the amount of genetic distance not accounted for by geographical distance to the degree of linguistic relatedness. While the paper is interesting, there is a series of problems altering its interpretation: first, the application of

192Which is not entirely appropriate in the case of distance matrices due to the non-independence of columns and rows, altering the significance level computed (Bonnet & Van de Peer, 2002:2).

simple Pearson's correlation coefficients to distance matrices is not appropriate (Footnote 192; Annex 4), and the same critique applies also to the linear regression of geographical on genetic distances[193]. Second, even if the residuals of the linear regression represent the amount of genetic variation not explained by geography, their binning using linguistic closeness is problematic.

The detection of boundaries and their comparison was also used: for example, Rosser *et al.* (2000), detected the boundaries (zones of sharpest genetic change) for their Y-chromosome data and tried to correlate them with linguistic differences. The map of the detected genetic boundaries is in Figure 38, where the thin lines represent the Delaunay connections[194] between the locations of the samples and the thick lines the genetic boundaries. The authors count the proportion of such genetic boundaries between different language families (64.2%), between subfamilies (40.5%) and within subfamilies (32.6%), but the differences between them are not significant (three-way χ^2), suggesting that "language may not be the primary force contributing to genetic barriers here" (Rosser *et al.*, 2000:1539):

> "[...] linguistic differences tend to cause some [i.e., slight] degree of population subdivision, regardless of whether such differences are between language families, between languages of the same family, or even between dialects of the same language" (Rosser *et al.*, 2000:1541).

193It would have been interesting to see the residuals plot in order to evaluate the heteroscedasticity and nonlinearity so that the appropriateness of linear regression could be assessed (Tabachnick & Fidell, 2001:116-122).

194The Delaunay triangulation (Fortin & Dale, 2005:60-62) is formed by joining all triplets of points for which the circle circumscribing the triangle formed by them does not contain any other points.

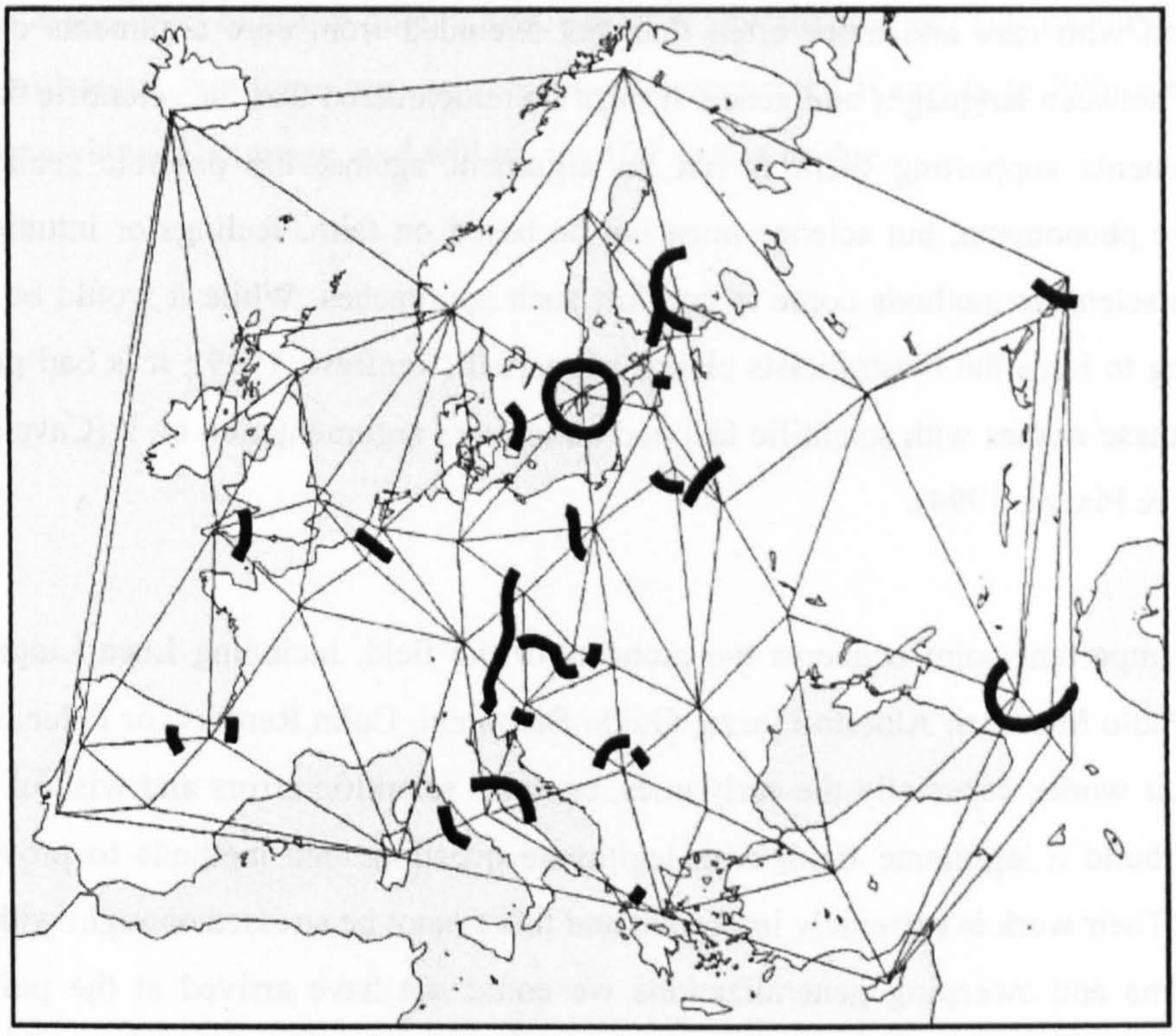

Figure 38: The genetic boundaries in the Y-chromosome distribution across Europe.

Reproduced from Rosser et al. (2000:1538, Figure 6). See text for details.

3.2.5. What do we know about languages and genes?

As the previous sections have shown, the study of the correlations between linguistic and genetic diversities is fraught with important difficulties, and the vast literature concerning the subject seems far from reaching a consensus, even on the most basic facts. A complicating factor, common to all the young and, for various reasons, fashionable, interdisciplinary fields, is represented by unsubstantiated claims, flawed methodologies, shallow understanding of one or more of the involved areas, or even lack of competence (such an extreme example, Arnaiz-Villena, Martińez-Laso & Alonso-Garciá (2001), is thoroughly analyzed in Annex 3). Nevertheless, what seems clear is that the relationship between the two is very complex, and that we must analyze regional cases one by one in all their complexity: demographic, socio-cultural and historical (McMahon, 2004).

Bold but linguistically problematic claims, like Nostratic, Eurasiatic, Amerind or the existence of a unitary pre-Indo-European circum-Mediterranean linguistic family, must be

considered with care and more often than not excluded from core arguments concerning parallels between languages and genes. It must be remembered that the scientific thinness of the arguments supporting them is not an argument against the possible reality of the respective phenomena, but science must not be based on faith, feelings or intuitions, until orthodox scientific methods come to comfort such approaches. While it would be certainly interesting to have the Nostraticists claims supported (Renfrew, 1999), it is bad practice to confuse these wishes with scientific fact and build one's argumentation on it (Cavalli-Sforza, Menozzi & Piazza, 1994).

Another important point concerns the pioneers of the field, including Luca Luigi Cavalli-Sforza, Paolo Menozzi, Alberto Piazza, Guido Barbujani, Colin Renfrew or Peter Bellwood: while their works, especially the early ones, contains scientific errors and wishful thinking, they *did* build a legitimate field, with legitimate questions and methods to provide valid answers. Their work is extremely important and this cannot be stressed enough: without their bold claims and sweeping generalizations we could not have arrived at the point where regional problems can be meaningfully asked and, probably, solved. And it is probably the fate of all pioneers to make such bold claims and sweeping generalizations.

We need better sampling strategies, better treatment of the linguistic data and integration with historical and socio-cultural variables. We also a need models explaining linguistic diversity, moving beyond simplistic mass migrations, towards a thorough understanding of the interaction between groups, in a metapopulation-like model. Dixon's "punctuated equilibrium" seems a good start but it is still far from satisfactory.

3.3. Conclusions: genes and language(s)

For the moment, there seems to be no bridging between these two types of relationships between genes and language(s): on one hand, there are the causal connections between inter-individual differences in genetic makeup and various linguistic aspects, potentially illuminating the structure, development, functioning and evolution of the "capacity for language", while, on the other hand, there are spurious correlations, due to common historical, cultural or demographic processes or events, between genetic and linguistic

diversities at the level of entire populations, possibly illuminating our (pre)history. The next chapters will argue that these two aspects *must* be connected if one is to fully embrace an evolutionary view of language, and will try to offer such a bridge.

4. A feature-based, spatial statistic approach to linguistic and genetic patterns

4.1. Introduction and hypotheses

The study of the correlations between languages and genes is still in its infancy (Chapter 3), without generally agreed-upon techniques, assumptions and standards. This Chapter will try, by building on a diverse literature, to introduce and adapt specific methods for studying the interactions between linguistic and genetic diversities in a geographical (spatial) and historical (temporal) context. These methods are inspired from classical and spatial statistics, geo-statistics, evolutionary genetics, linguistic typology and historical linguistics, and are tailored (where necessary) to the particularities of this specific field of research. One of the goals is the wider acceptance, testing and usage of this class of techniques in linguistic typology, studies of linguistic diversity and research concerning the interactions between genes and languages. By using these methods, novel and very interesting patterns and interactions, both linguistic and genetic, are uncovered, which warrant further study, potentially relevant to many disciplines, including evolutionary linguistics, typology, historical linguistics, areal linguistics, prehistory, humans evolution and psycholinguistics, etc.

But the main goal of this Chapter, in the context of this Thesis, is to statistically test, by appropriately using these techniques, the theory of *non-spurious correlations* between genetic and linguistic diversities, incarnated in a particular hypothesis, namely, that two specific human genetic haplogroups and one linguistic feature show strong and significant correlations, not entirely explainable by geographical or common descent (historical linguistic) processes. This correlation is highly significant in an inferential statistical sense, and also in the top of a vast sample of 983 genetic variants and 26 linguistic features, reducing very much the probability that it might be due to general processes shaping the relationship between genes and languages. As detailed in Section 4.9, this *a priori* hypothesis cannot be falsified with the available data, thus supporting the view of a correlation between this pair of genetic variants and the linguistic feature, but, as is well-known, statistics by itself cannot assess causal relationships. Nevertheless, it will be argued that this supports the general theory of non-spurious correlations, and a number of more

powerful tests is proposed, but a thorough discussion of its context, meaning and relevance will be postponed until the next Chapter (5).

Parts of the work forming the basis of this Chapter (especially the collection of some linguistic data) and invaluable in-depth discussions and insights are due to Prof. D. R. Ladd. All his contributions will be acknowledged as such in the text and, unless otherwise stated, all the remaining represent my original work.

4.2. The dataset: populations, genetic variants and linguistic features

In September 2005, two papers appeared in Science (Evans *et al.*, 2005; Mekel-Bobrov *et al.*, 2005), published by (almost) the same team, and dealing with two human genes involved in brain growth and development: *ASPM* (Abnormal Spindle-like, Microcephaly-associated; *MCPH5*, OMIM 605481, 1q31) and *Microcephalin* (*MCPH1*, OMIM 607117, 8p23; denoted as *MCPH* in the following). There seem to exist two main types of severe congenital microcephaly[195] (Gilbert, Dobyns & Lahn, 2005), "a 'high-functioning' group characterized by relatively mild phenotypes, and a 'low- functioning' group with much more severe phenotypes" (Gilbert, Dobyns & Lahn, 2005:585), and that those genes involved in high functioning congenital microcephalias (*ASPM*, *MCPH* and *SHH* - sonic hedgehog, OMIM 600725) show signatures of adaptive natural selection in the lineage leading to *Homo sapiens* (Gilbert, Dobyns & Lahn, 2005:585-586), suggesting that they might be involved in the human-specific patterns of brain growth and development (Gilbert, Dobyns & Lahn, 2005). Both *ASPM* and *MCPH* are involved in this high-functioning group of microcephalias (Gilbert, Dobyns & Lahn, 2005), as deleterious mutations in any of them determine a microcephalic phenotype (Mekel-Bobrov *et al.*, 2005:1720; Evans *et al.*, 2005:1717). There are to date six known loci associated with microcephalias (denoted *MCPH1-MCPH6* and including *MCPH*, *ASPM*, *CDK5RAP2* and *CENPJ*) (Evans *et al.*, 2005:1717).

For both *ASPM* and *MCPH*, a polymorphism was identified which shows signs of positive natural selection. These haplogroups were denoted *D* (for derived), giving thus a *D* haplogroup for *ASPM* (denoted *ASPM-D*) and a *D* haplogroup for *MCPH* (denoted *MCPH-*

195Defined as "a disorder that is characterized by marked reduction in brain size, with or without other abnormalities" (Gilbert, Dobyns & Lahn, 2005:581).

D), and defined as those groups of haplotypes containing the derived alleles G for polymorphism A44871G (*ASPM*, Mekel-Bobrov *et al.*, 2005:1721) and C for polymorphism G37995C (*MCPH*, Evans *et al.*, 2005:1718). The ages of these derived haplogroups was estimated at 5.8ky (0.5-14.1ky 95% CI) for *ASPM-D* (Mekel-Bobrov *et al.*, 2005:1721) and 37ky (14-60ky 95% CI) for *MCPH-D* (Evans *et al.*, 2005:1718; see Section 2.2.3 and Evans *et al.*, 2006 for details on its origin). Both *ASPM-D* and *MCPH-D* are geographically differentiated (Evans *et al.*, 2005:1718f; Mekel-Bobrov *et al.*, 2005:1721f). Both genes show accelerated evolution in the human lineage (~2 favorable changes/my), with *MCPH* evolving preponderantly during the early (MRCA of simians and great apes) and *ASPM* during the late (MRCA of great apes to humans) stages (Gilbert, Dobyns & Lahn, 2005:586). Thus, these two genes represent strong candidates for involvement in the evolution of human-specific traits, and, even if their exact function is not clear, they seem critical regulators of brain growth and development in humans. Their two recent *D* haplogroups are the first serious candidates of adaptive changes, geographically patterned, and not yet fixated factors involved in the brain growth and development of *Homo sapiens*.

Their geographic patterning, coupled with the signatures of natural selection and recent origin, point to ongoing evolution of the human brain (and, presumably, cognition) and their distribution across the human populations represents a snapshot of the process of fixation[196] (Mekel-Bobrov *et al.*, 2005; Evans *et al.*, 2005). For example, in the case of *ASPM-D*, the very strong geographic patterning can have multiple explanations:

> One is that haplogroup D first arose somewhere in Eurasia and is still in the process of spreading to other regions. The other is that it arose in sub-Saharan Africa, but reached higher frequency outside of Africa partly because of the bottleneck during human migration out of Africa. Finally, it is possible that differential selective pressure in different geographic regions is partly responsible (Mekel-Bobrov *et al.*, 2005:1722),

and for *MCPH-D*:

> Such population differentiation may reflect a Eurasian origin of haplogroup D, local adaptation, and/or demographic factors such a bottleneck associated with human migration out of Africa 50,000 to 100,000 years ago (Evans *et al.*, 2005:1719-1720),

but probably the most parsimonious (in both cases), is the first one. It must be stressed that these genes do not offer in any way support for racist ruminations[197]. This interpretation

---

196An alternative possibility is that the pattern is stable and represents the result of competing selective pressures.

197Unfortunately, but as expected, these two papers were immediately hijacked by individuals with

involving positive natural selection was contested by Currat *et al.* (2006), but the detailed response appearing in the same number (Mekel-Bobrov *et al.*, 2006) concludes that the best explanation for the patterning of *ASPM-D* and *MCPH-D* remains positive natural selection.

Shortly after the publication of these two papers (Evans *et al.*, 2005; Mekel-Bobrov *et al.*, 2005), prof. D. R. Ladd[198] and I[199] arrived at a hypothesis linking the frequency of *ASPM-D* and *MCPH-D* in a population and the linguistic usage of tone distinctions in the corresponding language(s). This hypothesis can be formulated as:

> There is a causal relationship between the frequency of *ASPM-D* and *MCPH-D* in a population and the probability that *tone* contrasts are used in the corresponding language(s),

and represents a case of non-spurious correlation between genetic and linguistic diversities. This hypothesis was based on apparently congruent geographical patterns and on a putative decomposition of the various linguistic strategies into sequential and parallel components (D. R. Ladd, *pc*), supported by data from linguistics and neurosciences. Unfortunately, for the moment, we do not have a coherent theory concerning the parallel and sequential mechanisms in language, and, subsequently, there is no clear mechanism linking *ASPM-D* and *MCPH-D* to tone.

This chapter describes a statistical approach to testing this hypothesis, the methods developed to tackle it and the results obtained so far. It is argued that not only these results are highly suggestive of an interesting link between *ASPM-D*, *MCPH-D* and tone and further studies using better samples and more advanced techniques are warranted, but also that the methodology developed is generally applicable to language-genes correlation as well as to other linguistic diversity studies.

4.2.1. The populations

The sampling used in this study is represented by the populations reported in Evans *et al.*

racist agendas and used as "scientific arguments" for their ideas. See, for example, Steve Sailer's http://www.vdare.com/Sailer/050911_new_orleans.htm and Annex 2.

198 With a long time interest in linguistic tone.

199 Looking for years for an example of non-spurious correlation between genetic and linguistic diversities.

(2005) and Mekel-Bobrov *et al.* (2005), which will be referred as the *E/MB sample*. They used a total of 59 worldwide populations, as follows (Mekel-Bobrov *et al.*, 2005:1721, caption of Fig. 1 and Evans *et al.*, 2005:1719, caption of Fig. 3):

Southeastern and Southwestern Bantu (South Africa), **San** (Namibia), **Mbuti Pygmy** (Democratic Republic of Congo), **Masai** (Tanzania), **Sandawe** (Tanzania), **Burunge** (Tanzania), **Turu** (Tanzania), **Northeastern Bantu** (Kenya), **Biaka Pygmy** (Central African Republic), **Zime** (Cameroon), **Bakola Pygmy** (Cameroon), **Bamoun** (Cameroon), **Yoruba** (Nigeria), **Mandenka** (Senegal), **Mozabite** [Algeria (Mzab region)], **Druze** [Israel (Carmel region)], **Palestinian** [Israel (Central)], **Bedouin** [Israel (Negev region)], **Hazara** (Pakistan), **Balochi** (Pakistan), **Pathan** (Pakistan), **Burusho** (Pakistan), **Makrani** (Pakistan), **Brahui** (Pakistan), **Kalash** (Pakistan), **Sindhi** (Pakistan), **Hezhen** (China), **Mongola** (China), **Daur** (China), **Orogen**[200] (China), **Miaozu** (China), **Yizu** (China), **Tujia** (China), **Han** (China), **Xibo** (China), **Uygur** (China), **Dai** (China), **Lahu** (China), **She** (China), **Naxi** (China), **Tu** (China), **Cambodian** (Cambodia), **Japanese** (Japan), **Yakut** [Russia (Siberia region)], **Papuan** (New Guinea), **NAN Melanesian** (Bougainville), **French Basque** (France), **French** (France), **Sardinian** (Italy), **North Italian** [Italy (Bergamo region)], **Tuscan** (Italy), **Orcadian** (Orkney Islands), **Russian** (Russia), **Adygei** [Russia (Caucasus region)], **Karitiana** (Brazil), **Surui** (Brazil), **Colombian** (Colombia), **Pima** (Mexico) and **Maya** (Mexico).

Table 6: The 59 world-wide populations in the E/MB sample.

***Bold**: the population's name; in parentheses, the population's geographic region/country.*

Unfortunately, there is no Australian sample. Also, the Americas are too poorly sampled, given their linguistic and genetic diversity, to be used in this study, so that the 5 American populations (Karitiana, Surui, Colombian, Pima and Maya) were excluded, but used as a test case. The resulting sample, composed of 54 populations, will be denoted as the *OWF sample* (Old World Full sample). Given the scarcity of information concerning the *OWF sample* in Evans *et al.* (2005) and Mekel-Bobrov *et al.* (2005), and the obvious ambiguity of some populations (e.g., Papuan, Southeastern and Southwestern Bantu), I have tried to refine this

200Probably a spelling mistake in the original papers, instead of **Oroqen**, but kept for ease of reference.

as much as possible, using information about the genetic samples taken from the ALFRED database (Rajeevan *et al.*, 2003; Osier *et al.*, 2002), linguistic information from the The Ethnologue (Gordon, 2005), and geographical and political information from the Maps of World (ref. Maps of World) and The World Factbook (ref. The World Factbook). Table 7 contains the identification information for each of the 59 populations of the *E/MB sample*[201], while Annex 5 contains more details.

The considered populations are very different in terms of political status, clarity of definition (Papuan to Burusho or Sindhi) and size (She, ~911 to Han, ~1 billion), highlighting again the inherent problems of non-systematic sampling (Section 3.2.4.3). Also in Annex 5, the usage of the amalgamated "Bantu speakers" sample to handle the missing data in sub-Saharan Africa is discussed. Overall, the most important aspect of this *E/MB sample* is its *reduced size and lack of systematicity*, so that any results are to be considered preliminary and in need of better sampling. Figure 39 shows the approximate geographical position of these 54 populations.

201Independently checked by D. R. Ladd.

ID	*Population full name*	*Population short name*	*Country/ region*	*Lng.*	*Linguistic family*	*Main city/ region*	*Geographical coordinates*
01	Southeastern and Southwestern Bantu	SESWBantu	South Africa		Niger-Congo	Pretoria	25°45′S/ 28°17′E
02	San	San	Namibia	naq	Khoisan	Windhoek	22°56′S/ 17°9′E
03	Mbuti Pygmy	Mbuti	Democratic Republic of Congo	efe	Nilo-Saharan	Bunia	1°34′N/ 30°15′E
04	Masai		Tanzania	mas	Nilo-Saharan	Arusha	3°22'S/ 36°40'E
05	Sandawe		Tanzania	sad	Khoisan	Dodoma	6°15'S/ 35°45'E
06	Burunge		Tanzania	bds	Afro-Asiatic	Kondoa	4°54'S/ 35°47'E
07	Turu	Turu	Tanzania	rim	Niger-Congo	Singida	6°00'S/ 34°30'E
08	Northeastern Bantu	Kikuyu	Kenya	kik	Niger-Congo	Nairobi	1°16′S/ 36°48′E
09	Biaka Pygmy	Biaka	Central African Republic	axk	Niger-Congo	Nola	3°35′N/ 16°04′E
10	Zime		Cameroon	lme	Afro-Asiatic	Garoua	9°19′N/ 13°21′E
11	Bakola Pygmy	Bakola	Cameroon	gyi	Niger-Congo	Kribi	2°57′N/ 9°56′E
12	Bamoun	Bamoun	Cameroon	bax	Niger-Congo	Foumban	5°45′N/ 10°50′E
13	Yoruba	Yoruba	Nigeria	yor	Niger-Congo	Ibadan	7°22′N/ 03°58′E
14	Mandenka	Mandenka	Senegal	mnk	Niger-Congo	Ziguinchor	12°34' N/ 16°16'W
15	Mozabíte	Mozabite	Algeria (Mzab)	mzb	Afro-Asiatic	Ghardaia	32°20′N/ 03°37′E
16	Druze	Druze	Israel (Carmel)	apc	Afro-Asiatic	Haifa	32°46′N/ 35°00′E
17	Palestinian	Palestinian	Israel (central)	ajp	Afro-Asiatic	Jerusalem	31°47′N/ 35°10′E
18	Bedouin	Bedouin	Israel (Negev)	ayl	Afro-Asiatic	Rahat	31°33′N/ 34°47′E

ID	*Population full name*	*Population short name*	*Country/ region*	*Lng.*	*Linguistic family*	*Main city/ region*	*Geographical coordinates*
19	Hazara	Hazara	Pakistan	haz	Indo-European	Quetta	31°15'N/ 66°55'E
20	Balochi	Balochi	Pakistan	bgp	Indo-European	Quetta	31°15'N/ 66°55'E
21	Pathan	Pathan	Pakistan	pst	Indo-European	Quetta	31°15'N/ 66°55'E
22	Burusho	Burusho	Pakistan	bsk	Burushaski	Balochistan	27°30'N/ 65°00'E
23	Makrani	Makrani	Pakistan	bcc	Indo-European	Gwadar	25°10'N/ 62°18'E
24	Brahui	Brahui	Pakistan	brh	Dravidian	Kalat	29°08'N/ 66°31'E
25	Kalash	Kalash	Pakistan	kls	Indo-European	Balanguru	35°44'N/ 71°46'E
26	Sindhi	Sindhi	Pakistan	snd	Indo-European	Karachi	24°53'N/ 67°00'E
27	Hezhen	Hezhen	China	gld	Altaic	Harbin	45°48'N/ 126°40'E
28	Mongola	Mongola	China	mvf	Altaic	Hohhot	40°52'N/ 111°40'E
29	Daur	Daur	China	dta	Altaic	Nirji	48°28'N/ 124°28'E
30	Orogen[202]	Orogen	China	orh	Altaic	Alihe	50°34'N/ 123°43'E
31	Miaozu	Miaozu	China	hmy	Hmong-Mien	Guizhou	27°00'N/ 107°00'E
32	Yizu	Yizu	China	yif	Sino-Tibetan	Minjian	28°50'N/ 103°32'E
33	Tujia	Tujia	China	tji	Sino-Tibetan	Jishou	28°19'N/ 109°43'E
34	Han	Han	China	cmn	Sino-Tibetan	Beijing	39°55'N/ 116°20'E
35	Xibo	Xibo	China	sjo	Altaic	Shenyang	41°48'N/ 123°27'E
36	Uygur	Uygur	China	uig	Altaic	Urumqi	43°43'N/ 87°38'E

202See footnote 200.

ID	*Population full name*	*Population short name*	*Country/ region*	*Lng.*	*Linguistic family*	*Main city/ region*	*Geographical coordinates*
37	Dai	Dai	China	tdd	Tai-Kadai	Jinghong	21°27'N/ 100°25'E
38	Lahu	Lahu	China	lhu	Sino-Tibetan	Kunming	25°01'N/ 102°41'E
39	She	She	China	shx	Hmong-Mien	Fuzhou	26°05'N/ 119°16'E
40	Naxi	Naxi	China	nbf	Sino-Tibetan	Lijiang	23°12'N/ 108°9'E
41	Tu	Tu	China	mjg	Altaic	Xining	36°34'N/ 101°40'E
42	Cambodian	Cambodian	Cambodia	khm	Austro-Asiatic	Phnom Penh	11°33'N/ 104°55'E
43	Japanese	Japanese	Japan	jpn	Japanese	Tokyo	35°41'N/ 139°46'E
44	Yakut	Yakut	Russia (Siberia)	sah	Altaic	Yakutsk	62°12'N/ 129°44'E
45	Papuan		New Guinea			New Guinea Island	05°00'S/ 140°00'E
46	NAN Melanesian	NANMelanesian	Bougainville	nas	East Papuan	Bougainville	06°00'S/ 155°00'E
47	French Basque	FrBasque	France	eus	Basque	Bayonne	43°30'N/ 01°28'W
48	French	French	France	fra	Indo-European	Paris	48°50'N/ 02°20'E
49	Sardinian	Sardinian	Italy	src	Indo-European	Cagliari	39°13'N/ 09°07'E
50	North Italian	NItalian	Italy (Bergamo)	vec	Indo-European	Bergamo	45°41'N/ 09°43'E
51	Tuscan	Tuscan	Italy	ita	Indo-European	Firenze	43°47'N/ 11°15'E
52	Orcadian	Orcadian	Orkney Islands	sco	Indo-European	Kirkwall	59°09'N/ 02°59'W
53	Russian	Russian	Russia	rus	Indo-European	Moskva	55°45'N/ 37°37'E
54	Adygei	Adygei	Russia (Caucasus)	ady	North Caucasian	Maykop	44°36'N/ 40°05'E
55	Karitiana		Brazil	ktn	Tupi	Porto Velho	08°46'S/ 63°54'W

ID	Population full name	Population short name	Country/ region	Lng.	Linguistic family	Main city/ region	Geographical coordinates
56	Surui		Brazil	sru	Tupi	Vilhena	12°40'S/ 60°05'W
57	Colombian		Colombia	pio	Arawakan	Puerto Carreno	06°12'N/ 67°22' W
58	Pima		Mexico	pia	Uto-Aztecan	Hermosillo	29°10' N/ 111°00' W
59	Maya		Mexico	yua	Mayan	Mérida	20°58' N/ 89°37' W

Table 7: Geographic/politic and linguistic information for the 59 populations in the E/MB sample.

ID: the population's numeric identification used in the E/MB sample, Population full name: the population's name in the E/MB sample, Population short name: he population's name as used in the present work, Country/region: the population's geographical information as reported in the E/MB sample, Lng.: the 3 letter language code as defined in Gordon (2005), Linguistic family: the historical linguistic affiliation, as given by Gordon (2005), Main city/region: the capital or most important city of the region where the language is reportedly spoken (or, if none, the entire region is reported), and Geographical coordinates: the "Main city/region"'s geographical coordinates. The populations without a short name have not been included in the final sample.

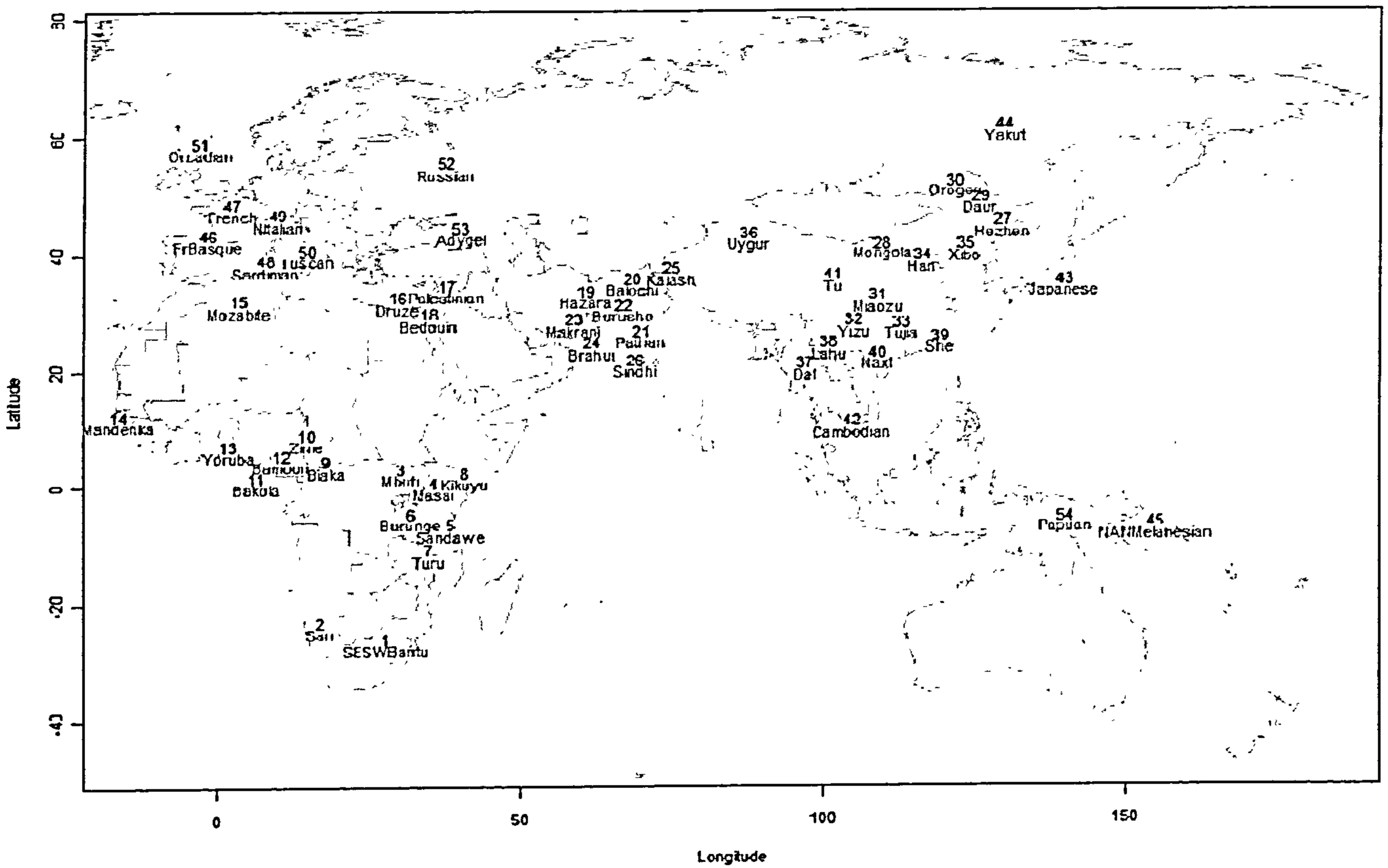

Figure 39: The approximate geographical positions of the 54 populations in the *OWF* sample.

Black: ID, gray : population name. The positions have been adjusted to fit.

4.2.2. The genetic data

For the 54 populations of the *OWF sample*, information concerning the frequency of various genetic variants was gathered from three sources:

- The *two original papers*: the frequency of *ASPM-D* from Evans *et al.* (2005:1721, Fig. 1's caption) and of *MCPH-D* from Mekel-Bobrov *et al.* (2005:1719, Fig. 3's caption);
- The *Allele Frequency Database (ALFRED)*: this online database (Rajeevan *et al.*, 2003; Osier *et al.*, 2002) provides allele frequency information for many loci and populations relevant for anthropological, human evolutionary or medical studies. An initial number of 133 genetic variants was selected from the entire set of 10195 available, covering the largest possible set of populations from the *OWF sample* (the only selection criterion was to cover at least 44 populations);
- The *Human Diversity Panel Genotypes* (*HDGP*) database: this represents the dataset used by Rosenberg *et al.* (2002). An initial number of 1029 genetic variants was selected from the entire available set, covering the largest possible set of populations from the *OWF sample* (the only selection criterion was to cover at least 44 populations).

For each of these 1164 selected genetic variants, various non-frequency information was also gathered:

- the *locus*: this is the name of the genetic locus as used in the original source;
- the *full name*, *alternate name* and *site*: full name(s) of the locus – the actual denomination of this information and number of names depends on source and locus;
- the *allele*: the actual allele at the locus considered. This information generally represents the number of repeats in *Short Tandem Repeat Polymorphisms*[203] (*STR*s), as these represent the huge majority of these markers (all except *ASPM* and *MCPH*)[204];

---

203 Short patterns of nucleotides which is repeated in sequence a variable number of times; also known as *microsatellites* (Jobling, Hurles & Tyler-Smith, 2004:64-68).

204 An anonymous reviewer of an earlier version suggested that also SNP (Single Nucleotide Polymorphisms) should be included. But, while this observation is *in principle* valid, practically no SNP was found in the databases to comply with the criteria. Also, from a purely theoretical point of view, I cannot see any relevant differences between SNPs and STRs which could bias this

- the *chromosome*: this represents the chromosome the markers maps onto (1-22, X and Y);
- the *physical position on the chromosome*: the information concerning their physical position on the chromosome and was gathered using the *UniSTS Project*[205], as the *Marshfield position* (measured in cM), physical map position as reported by both the Human Genome Project (*human genome project physical position*, measured in bp) and the Celera Genomics project (*Celera project physical position*, measured in bp).

Many sub-Saharan African populations (9 out of 14: SESWBantu, Masai, Sandawe, Burunge, Turu, Kikuyu, Zime, Bakola and Bamoun) systematically lack such data, making them unusable in the analyses. Therefore, using a missing data handling procedure (Annex 5; Tabachnick & Fidell, 2001:58-66), the "Bantu speakers" amalgamated sample was used to provide frequency data for 5 Bantu-speaking populations (SESWBantu, Turu, Kikuyu, Bakola and Bamoun)[206], reducing the number of populations to be deleted to 4 (Masai, Sandawe, Burunge and Zime).

4.2.3. The linguistic data

This is composed of two parts: the languages assigned to the 54 populations of the *OWF sample*, and a set of linguistic features with their appropriate values in these languages. For the first part, it was assumed that for each population there is a single representative language spoken as the population's first (native) language. Such an assumption could be criticized as over-simplistic, but this critique might not be entirely justified, as most populations are defined using linguistic criteria in the original sources, and the values of the linguistic features were not rigidly gathered. Nevertheless, in some cases, such a simple relationship between populations and languages does not hold. The attribution of languages to populations is based mainly on information in the Ethnologue (Gordon, 2005) and is summarized in Table 7 above, and further detailed in Annex 5. Most such attributions are fairly straightforward and uncontroversial, but some must be discussed[207]:

- *SESWBantu*: no single language was attributed, but data from Xhosa (xho) and Zulu

kind of analysis.

205 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unists (September, 2006).

206 The original frequencies of *ASPM* and *MCPH* were kept.

207 This linguistic attribution was checked for consistency with prof. D. R. Ladd.

(zul) was preferentially used;

- *San*: Nama (naq) was chosen to represent this population linguistically, because of the number of speakers and availability of information;
- *NANMelanesian*: as described in Annex 5, its linguistic attribution was far from obvious, but it turned out to be the Naasioi, [nas];
- *Papuan*: It is impossible (and unwarranted) to assume a single linguistic attribution for this sample: therefore, majority judgments at the scale of Papua-New Guinea were used.

For each language, its linguistic family was recorded, using a mostly uncontroversial classification, based on Gordon (2005), and avoiding any hazardous claims involving macrofamilies or, for example, the inclusion of Japanese in Altaic.

The second type of linguistic information is represented by linguistic features. A *linguistic feature* (*linguistic variable* or *typological parameter*) will be defined following the usage in Haspelmath, Dryer, Gil & Comrie (2005):

> a *structural property of languages* with respect to which in principle all languages can be defined, e.g. "tone", "clausal word order". The actual type that a language represents ("no tone", "tone"; "SVO", "SOV", etc.) is called a *value*. [...] For the purposes of this program, a feature can be taken to be a list of values
>
> (WALS Software, Help: Glossary: "feature", *italics* mine),

but similar concepts are developed and used by Croft (1990:27-39) and Comrie (1981:30-39). Linguistic features are traditionally used in typological linguistic classifications, the study of linguistic universals and implicational hierarchies (Croft, 1990; Comrie, 1981). The 141 linguistic features database in Haspelmath, Dryer, Gil & Comrie (2005) was carefully screened and 24 linguistic features were retained, which met the selection criteria:

- as good as possible coverage of the languages corresponding to *OWF sample*;
- meaningful collapsing of the values range into a binary classification.

While the requirement for the best covering of the considered languages (populations) is obvious, the second condition might seem artificial. Nevertheless, it is justified by the restrictions inherent in the statistical theory of measurement and its influence on the range of statistical tests applicable (de Vaus, 2002:40-46; Howitt & Cramer, 2003:5-8). If all linguistic features considered are uniformly binary, then the results can be meaningfully compared across features and, very importantly, most multivariate and spatial statistical methods treat binary variables as interval (Tabachnick & Fidell, 2001:112).

Besides these 24 binary linguistic features retained from Haspelmath, Dryer, Gil & Comrie (2005), another 2 binary and 2 numeric (interval) linguistic features were added by prof. D. R. Ladd, resulting in a gross total of 28 linguistic features. The details for each of these are given in Annex 6.1, while Annex 6.2 lists their values for each population. The names and short descriptions on these features are given in Table 8.

Linguistic feature	*Description*
ConsCat	The richness of consonant inventory.
Cons*	The actual number of consonants.
VowelsCat	The richness of vowel inventory.
Vowels*	The actual number of vowels.
UvularC	Are there uvular consonants?
GlotC	Are there glottalized consonants?
VelarNasal	Are there velar nasals?
FrontRdV	Are there front rounded vowels?
Codas	Are codas allowed?
OnsetClust	Are onset clusters allowed?
WALSSylStr	The complexity of syllable structure.
Tone	Does the language have a tonal system?
RareC	Does the language have any rare consonants?
Affixation	How much affixation does the language use?
CaseAffixes	Are cases marked with affixes?
NumClassifiers	Does the language have numeral classifiers?
TenseAspect	Are there tense-aspect marking inflections?
MorphImpv	Are there dedicated morphological categories for second person imperatives?
SVWO	What is the dominant Subject-Verb word order (if any)?
OVWO	What is the dominant Object-Verb word order (if any)?
AdposNP	What is the dominant order (if any) between adposition and noun phrase?
GenNoun	What is the dominant order (if any) between genitive and noun?
AdjNoun	What is the dominant order (if any) between adjective and noun?
NumNoun	What is the dominant order (if any) between numeral and noun?
InterrPhr	Is the interrogative phrase initial?
Passive	Is there a passive construction?
NomLoc	Are the encoding strategies for locational and nominal predications identical?
ZeroCopula	Is the omission of copula allowed?

Table 8: Summary listing of the 28 considered linguistic features

Details in Annexes 6.1 and 6.2; the starred (*) linguistic features are not binary.

4.3. Notes on data analysis

In general, statistical textbooks recognize three types of statistical techniques (e.g., Howitt & Cramer, 2003:3-4; Tabachnick & Fidell 2001:7-8):

- *descriptive statistics*: dedicated to summarizing the data at hand;
- *inferential statistics*: concerned with the confidence of generalizations from samples to populations, and
- *exploratory techniques*: designed to help the researcher to make sense of large amounts of data, also known as *data simplification*, *data mining*, *data reduction* or *data exploration*.

The analyses performed in this chapter cover all these aspects: first, the extensive linguistic and genetic sample gathered needs describing, using descriptive techniques. Second, specific hypotheses concerning the relationship between linguistic features, genetic variants and geographic structure, are formulated and tested, using inferential statistics. And third, exploratory techniques[208] are employed to suggest trends otherwise buried in the complexity of the data, so that more refined hypotheses can be formulated and appropriately tested.

Given the fact that, sometimes, many statistical (inferential) tests are performed using the same sample, appropriate *corrections for multiple comparisons* must be used to amend the probability that just by pure chance a significant result is obtained. Holm's (1979) multiple comparisons correction, which deals with this increase in probability of a *Type I error*[209] in multiple tests, is a modification of the *classical Bonferroni correction* (Wright, 1992:1008-1009; Walsh, 2004:4; Shaffer, 1995:569-570). In the classical Bonferroni correction, the overall (family-wise) p-level, p_{FW}, determines the comparison-wise p-level, p_{CW}, through the equation:

$$p_{CW} = p_{FW} / n$$

where n is the number of comparisons, but this correction is excessively conservative, increasing the *Type II error*[210] probability to unacceptable levels (Wright, 1992:1008; Walsh, 2004:3; Schaffer, 1995:569). Holm's multistage method is a *sequentially rejective*

208It must be pointed out that such techniques are enormously useful, but that their usage for valid inference must be appropriately controlled, using specific methods (e.g., multiple comparisons corrections, etc.).

209Type I error is the error of rejecting the null hypothesis when it is, in fact, true.

210Type II error is the error of accepting the null hypothesis when it is, in fact, false.

Bonferroni procedure which builds on this and which sequentially considers all the hypotheses, starting with the most improbable (smallest unadjusted p-value), and rejects them until the first one which cannot be rejected (for details see Wright, 1992:1008; Walsh, 2004:4; Schaffer, 1995:569). The generic term

$$p_i' = p_{CW} / (n-i+1)$$

represents the *adjusted p-value* of the hypothesis H_i. Given that Holm's method is much more powerful than Bonferroni's and does not have any limiting assumptions [as opposed to more powerful methods, like Hommel's (1988) or Hochberg's (1988)], it will be systematically used in these analyses[211] (denoted as *Holm mcc*). For good reviews of the very complex problem of multiple comparisons corrections, see, for example, Wright (1992) and Schaffer (1995). A subtle but very important observation is that *a priori formulated hypotheses do not need to be corrected for multiple comparisons*.

Another important observation concerns the interpretation of our extensive sample of linguistic and genetic data. The most important hypothesis sought to be tested in this analysis concerns the relationship between *ASPM*, *MCPH* and *tone*. The standard approach is to employ inferential statistical techniques to test this hypothesis and reject, or fail to reject it using a generally accepted significance level (e.g., $p < 0.05$), and, as will be shown below, the hypothesis of a non-null relationship between them cannot be rejected at such a significance level. But there is no *a priori* reason to assume that in this specific case, concerning the relationship between genetic and linguistic diversities, the correct null hypothesis is indeed the lack of any relationship. Potential arguments against assuming this null hypothesis are many, including previous claims in the literature (Section 3.2) and considerations of human history and prehistory. Thus, besides using standard inferential techniques, it is necessary to try to establish the behavior of as many as possible genetic variants and linguistic features, so that a baseline is provided, against which the hypothesized relationship can be tested. Therefore, when a certain *p*-level is inferential in standard statistical terms, it will be used as such, but when reference is made to the comparison against the entire sample, as, for example, the proportion of correlations *in the empirical sample* greater than a given correlation, without any generalization to the entire population of such linguistic features and genetic variants, the subscript "$_{sample}$", "level *x*" (instead of *p*) or expression "*level x in the sample*"[212], will be used. For such intra-sample comparisons, no

211As implemented by R's (R Development Core Team, 2006) `p.adjust` method.
212For example, "level 0.05 (two-tailed) in the sample" means "in the top 5% (two-tailed) of the

multiple comparisons correction is needed, as no inference takes place. It is assumed that if the considered relationship falls in the 5% tail (single or two-tailed, depending on the specific comparison) *relative to the empirical sample*, it is different enough from the "average" behavior of such relationships, so that its properties are deemed interesting.

All statistical analyses presented in this thesis used *The R Project for Statistical Computing* (http://www.r-project.org/) (R Development Core Team, 2006), version 2.3.1, a free and very powerful "software environment for statistical computing and graphics", and some of its packages, including *relimp*, *Design*, *maps*, *TeachingDemos*, *vegan*, *plotrix*, *tripack*, *nnet*, *car* and *sna*.

4.4. Analyzing the linguistic data

The original linguistic data (Annex 6.2) has the following characteristics:

- *number of populations*: 54;
- *number of linguistic features*: 28;
- *total number of missing data*: 94 (6.22%);
- *number of missing data across populations*: (Table 9) Most of the populations (26, 48.15%) have no missing data, while 7 (12.96%) have more than 20% (5) missing data and even if 6 out of these are in Asia, their geographical distribution does not suggest any systematicity. Also, their global distribution does not seem to follow any systematic pattern (Figure 40);
- *number of missing data across linguistic features*: (Table 10) Only 1 (3.57%) linguistic feature has more than 20% (12) missing data.

Population	*Number of missing data*	*Percent of missing data*
Bamoun	0	0.00%
Bedouin	0	0.00%
Brahui	0	0.00%
Burunge	0	0.00%
Cambodian	0	0.00%

empirical distribution".

Population	*Number of missing data*	*Percent of missing data*
Druze	0	0.00%
FrBasque	0	0.00%
French	0	0.00%
Han	0	0.00%
Hazara	0	0.00%
Japanese	0	0.00%
Mandenka	0	0.00%
Masai	0	0.00%
Mongola	0	0.00%
Naxi	0	0.00%
Orcadian	0	0.00%
Palestinian	0	0.00%
Pathan	0	0.00%
Russian	0	0.00%
San	0	0.00%
SESWBantu	0	0.00%
Sindhi	0	0.00%
Tujia	0	0.00%
Turu	0	0.00%
Yizu	0	0.00%
Yoruba	0	0.00%
Balochi	1	3.57%
Biaka	1	3.57%
Kalash	1	3.57%
Makrani	1	3.57%
Mozabite	1	3.57%
Sandawe	1	3.57%
Sardinian	1	3.57%
She	1	3.57%
Tuscan	1	3.57%
Burusho	2	7.14%
Dai	2	7.14%
Mbuti	2	7.14%
NItalian	2	7.14%
Orogen	2	7.14%
Zime	2	7.14%

Population	*Number of missing data*	*Percent of missing data*
Bakola	3	10.71%
Kikuyu	3	10.71%
NANMelanesian	4	14.29%
Papuan	4	14.29%
Uygur	4	14.29%
Miaozu	5	17.86%
Adygei	7	25.00%
Daur	7	25.00%
Lahu	7	25.00%
Yakut	7	25.00%
Hezhen	9	32.14%
Tu	9	32.14%
Xibo	9	32.14%

Table 9: The missing data analysis for populations.

Light gray: more than the recommended upper limit of 20% missing data.

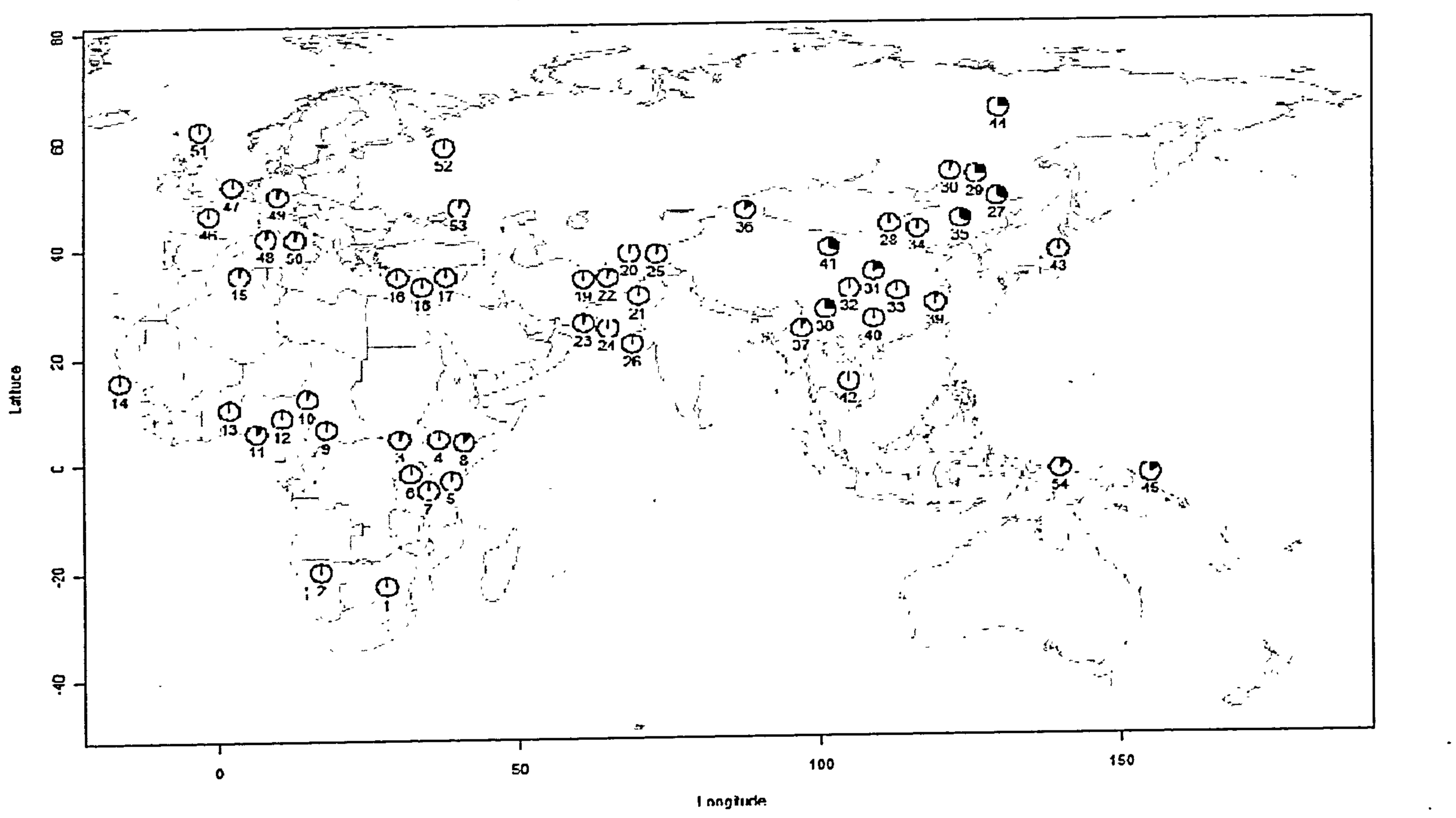

Figure 40: The map of linguistic missing data across languages.
Each pie chart represents the percent of missing data.

Linguistic feature	*Number of missing data*	*Percent of missing data*
ConsCat	0	0.00%
GlotC	0	0.00%
RareC	0	0.00%
Tone	0	0.00%
UvularC	0	0.00%
VowelsCat	0	0.00%
AdposNP	1	1.85%
Affixation	1	1.85%
CaseAffixes	1	1.85%
FrontRdV	1	1.85%
OVWO	1	1.85%
TenseAspect	2	3.70%
VelarNasal	2	3.70%
WALSSylStr	2	3.70%
AdjNoun	3	5.56%
Cons	3	5.56%
Passive	3	5.56%
SVWO	3	5.56%
Vowels	3	5.56%
NumNoun	4	7.41%
GenNoun	5	9.26%
ZeroCopula	5	9.26%
InterrPhr	6	11.11%
MorphImpv	8	14.81%
NumClassifiers	8	14.81%
NomLoc	10	18.52%
OnsetClust	10	18.52%
Codas	12	22.22%

Table 10: The missing data analysis for linguistic features.

Light gray: more than the recommended upper limit of 20% missing data.

With the exclusion of Papuan (see below), the distribution of the 28 linguistic features is given in the following Table (11) and boxplots (Figure 41):

Linguistic feature	*Percent 0s*
GenNoun	50.00%
AdposNP	51.92%
OVWO	51.92%
VowelsCat	52.83%
WALSSylStr	52.94%
Tone	54.72%
VelarNasal	44.23%
ZeroCopula	56.25%
CaseAffixes	59.62%
NomLoc	39.53%
NumNoun	61.22%
Codas	38.10%
AdjNoun	38.00%
ConsCat	62.26%
UvularC	62.26%
InterrPhr	63.83%
OnsetClust	65.12%
MorphImpv	31.11%
Passive	26.00%
RareC	77.36%
Affixation	21.50%
GlotC	79.25%
NumClassifiers	80.00%
TenseAspect	17.65%
FrontRdV	86.54%
SVWO	96.00%

Table 11: The distribution of values (0 and 1) for the 28 linguistic features in the 53 populations of the OWNP (OWF without Papuan) sample.

Values are percentages, sorted by the degree of deviation from the ideal 50%:50% distribution; gray cells represent linguistic features which are markedly skewed.

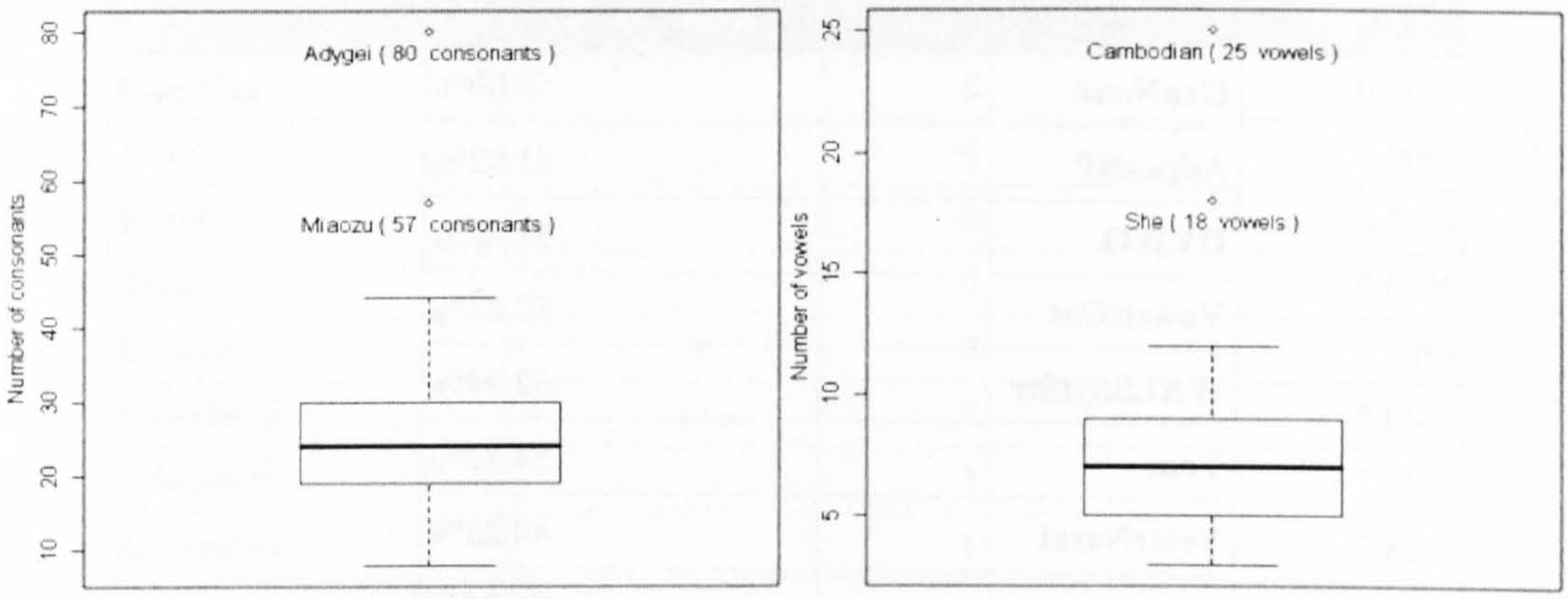

Figure 41: Boxplots of the actual number of consonants (Cons) and vowels (Vowels).
The outliers are: Adygey (80) and Miaozu (57) for consonants and Cambodian (25) and She (18) for vowels. The means (and medians) are 26.16 (24) consonants and 7.373 (7) vowels.

The actual number of vowels (*Vowels*) and consonants (*Cons*) correlate very well with their binary counterparts (*VowelsCat* and *ConsCat*, respectively) (Table 12). Therefore, only the 26 binary linguistic features ("binary" will be assumed by default) will be used in order to insure the comparability across the features.

Case	*t*	*df*	p_t	μ_0	μ_1	*r*	p_r
Consonants	-4.71	20.37	0.0001	20.50	35.68	0.64	0.0000
Vowels	-4.20	29.74	0.0002	5.57	9.57	0.54	0.0000

Table 12: The strong correlation between two measures of the vowel and consonant inventories.

The actual number of vowels/consonants vs the classification of the complexity of the vowel/consonantal systems. Two-samples t-tests and Pearson's r are significant at the 0.01 level. t: the value of the two samples t-test; df: the degrees of freedom; p_t:significance level of the t-test; μ_0, μ_1:the means of the two groups, corresponding to value 0 and 1 of the binary feature; r: the Pearson's correlation coefficient; p_r:its significance level.

Pearson's *r* correlation coefficients between all pairs of binary features[213] have mean = 0.012 and a sd = 0.274 (Figure 42). The 23 correlations significant at the .05 level (Holm mcc) are listed in Table 13.

213Equivalent to the *Phi* correlation coefficient for binary variables.

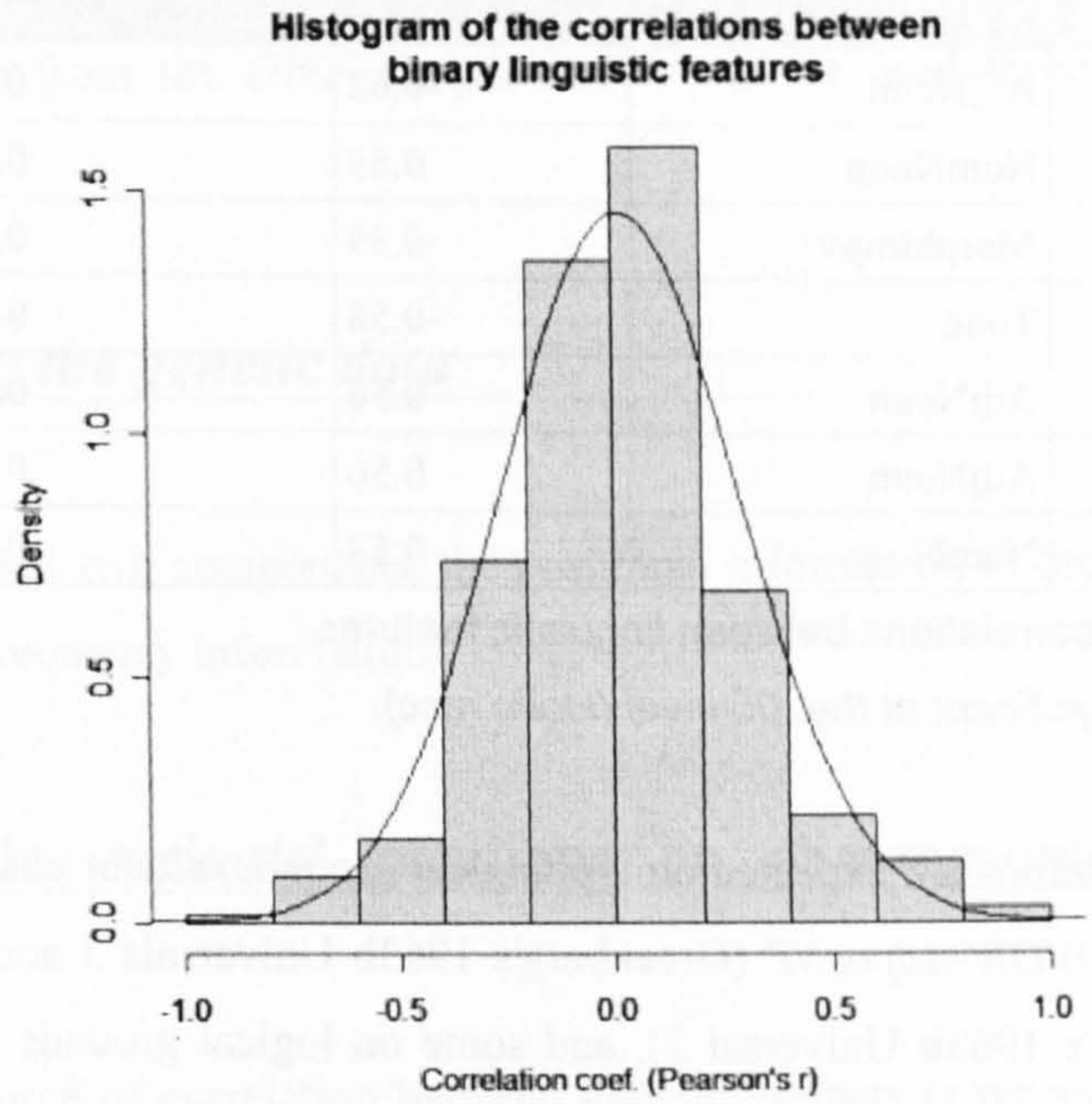

Figure 42: Histogram of the distribution of correlation coefficients (Pearson's *r*) between all pairs of linguistic features.

It approximates very well a normal distribution with mean 0.01209 and sd 0.2743.

1st linguistic feature	*2nd linguistic feature*	*Pearson's r*	*abs(Pearson's r)*	*adjusted p*
OVWO	AdposNP	0.96	0.96	0.0000
Affixation	TenseAspect	0.81	0.81	0.0000
Affixation	NumClassifiers	-0.80	0.80	0.0000
Codas	WALSSylStr	0.78	0.78	0.0000
TenseAspect	MorphImpv	0.74	0.74	0.0000
WALSSylStr	Tone	-0.73	0.73	0.0000
Tone	NumNoun	0.73	0.73	0.0000
Codas	Tone	-0.72	0.72	0.0000
NumClassifiers	TenseAspect	-0.69	0.69	0.0000
GenNoun	AdjNoun	0.68	0.68	0.0000
Affixation	MorphImpv	0.68	0.68	0.0000
AdposNP	GenNoun	0.66	0.66	0.0000
WALSSylStr	NumNoun	-0.66	0.66	0.0000
Codas	NumNoun	-0.65	0.65	0.0000
OnsetClust	WALSSylStr	0.65	0.65	0.0000
OVWO	GenNoun	0.63	0.63	0.0000

1st linguistic feature	*2nd linguistic feature*	*Pearson's r*	*abs(Pearson's r)*	*adjusted p*
CaseAffixes	AdjNoun	-0.62	0.62	0.0000
AdjNoun	NumNoun	0.59	0.59	0.0000
NumClassifiers	MorphImpv	-0.59	0.59	0.0000
OnsetClust	Tone	-0.58	0.58	0.0000
OVWO	AdjNoun	0.56	0.56	0.0000
AdposNP	AdjNoun	0.56	0.56	0.0000
GlotC	NumNoun	0.53	0.53	0.0001

Table 13: The correlations between linguistic features.

Pearson's r, significant at the .05 level (Holm mcc).

Some of these correlations are expected on typological grounds, as, for example, word order correlations[214], like *OVOW-AdposNP* (Greenberg's 1963b Universals 3 and 4) or *GenNoun-AdjNoun* (Greenberg's 1963b Universal 2), and some on logical grounds (i.e., the way the features are defined, like *Codas-WALSSylStr*), but many are not. All correlations are important, including the very interesting ones involving *Tone* vs. *WALSSylStr*, *NumNoun*, *Codas* and *OnsetClust*. Given the definition of *Codas*, *OnsetClust* and *WALSSylStr*, there is a high correlation between them, and we can consider that there is a subtending factor called *syllable structure*. Thus, there is a strong correlations between *Tone*, *syllable structure*, and *NumNoun*. This type of correlational table should be refined and its study seems very promising for typological research.

For the Papuan sample it is impossible to establish a unique value of *Tone* (Section 4.2.3), so that three cases must be considered:

- a value of 0 for *Tone*;
- a value of 1 for *Tone*;
- a case without the Papuan sample.

To asses the possible impact these different coding schemes can have, *t-tests* between the inter-linguistic features correlations were performed, and they proved to be non-significant ($p > 0.95$ in all cases), showing that the inclusion of the Papuan sample as either tonal or non-tonal (or its exclusion) seems not to affect the results. Moreover, from a genetic point of view, this sample is highly problematic (Section 4.9), so that its exclusion is highly

214There are many functional explanations proposed for these universals, see, for example, Kirby 1999.

recommended. Given these, a new *OWNP sample* (*Old World No Papuan*) was created by excluding Papuan from the *OWF sample* and it will be used throughout the following analyses.

4.5. Analyzing the genetic data

The genetic data has two components, the positional information of the genetic variants and their population frequency information.

4.5.1. Genetic variants' positions on chromosomes and genetic linkage

One potential source of correlation between genetic variants is represented by the *genetic linkage* between them, but preliminary analyses seem to suggest that this represents a rather minor issue for our data. From the initial 1164 genetic variants, after deleting those markers for which there was no information concerning the chromosome they are on (2) or no positioning information (48), there remained a total of 1114 markers, distributed across chromosomes as shown in Figure 43.

4.5.2. The genetic variants' frequencies in populations

The genetic variants duplicated between the databases were identified using their full names(s), position and specific allele, and 124 pairs resulted. For each pair, the variant covering most of the population was retained: systematically, the *HDGP* database proved richer than *ALFRED* by one or two populations (*Makrani* ánd *NItalian*). Moreover, 9 genetic variants were also deleted, as they introduced systematic missing data in Africa, resulting a final list of 981 genetic variants.

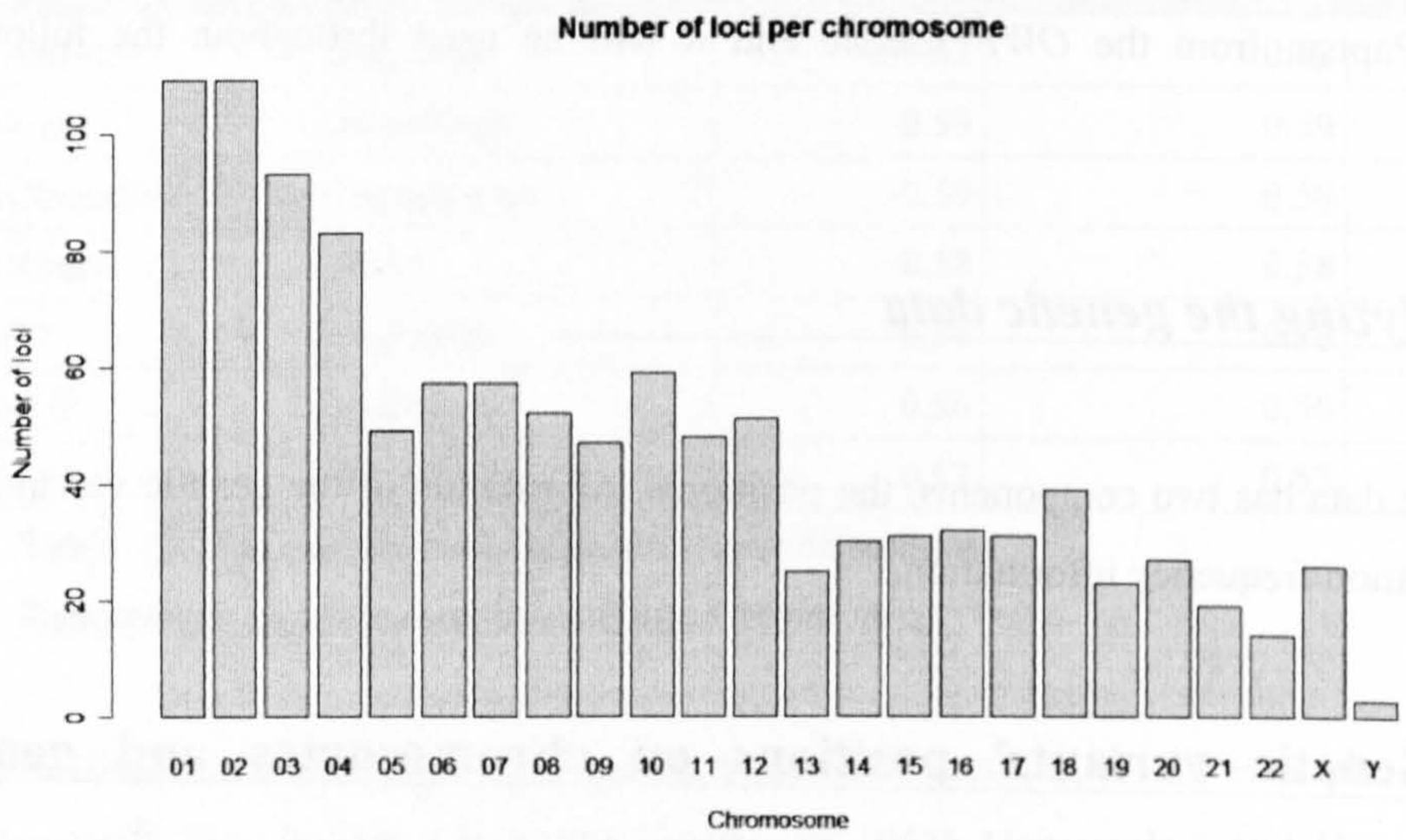

Figure 43: The number of alleles per chromosome.

The chromosomes are numbered from 1 to 22 in decreasing size order . The Y chromosome is very poor in coding DNA (Jobling, Hurles & Tyler-Smith, 2004:33-34).

As previously discussed (Sections 4.2.1 and 4.2.2), because there were systematic and almost total missing data in a number of African populations, data from the amalgamated "Bantu speakers" sample was used for those samples belonging to the "Bantu" linguistic group [Gordon's (2005) Niger-Congo: Bantoid: Narrow Bantu]: *SESWBantu*, *Turu*, *Kikuyu*, *Bakola* and *Bamoun*. This approach to missing data is assumed to introduce minimal distortion as, plausibly, the Bantu populations are the result of fairly recent demographic expansions (Cavalli-Sforza, Menozzi & Piazza, 1994:158-194, esp. 162-163; Jobling, Hurles & Tyler-Smith, 2004:324-328). Nevertheless, a bias towards those linguistic features uniform across the sampled Bantu languages and against those showing variation is introduced. Therefore, better genetic data for sub-Saharan Africa is needed before any definitive conclusions can be reached. To assess the effects of this missing data handling procedure, two artificial genetic variants, *ASPM** and *MCPH**, were created from *ASPM* and *MCPH*, respectively, by replacing their actual frequency values in the 5 Bantu populations with their average. Both the original and the artificially created genetic variants were used during the following analyses and the overall results clearly support the view that their behavior is essentially the same[215], suggesting that this procedure does not distort the results too much.

215E.g,, the correlations between all pairs of genetic variants have mean = 0.02373297 and sd = 0.2249763, originally, and mean = 0.02395730 and sd = 0.2254099 after *ASPM** and *MCPH** are included.

Unfortunately, even this procedure cannot cope with the systematic and almost total missing data for *Masai*, *Sandawe*, *Burunge* and *Zime*, which were, therefore, deleted, resulting in the *OWG sample*[216]. The resulting database has the following characteristics:

- *number of populations*: 50;
- *number of genetic variants*: 983;
- *total number of missing data*: 610 (1.24%);
- *number of missing data across populations*: (Table 14) No population exceeds the recommended 20% upper limit and the global distribution of missing data across populations does not seem to follow any systematic pattern;
- *number of missing data across genetic variants*: The maximum missing data is 2 (4%) for 166 markers.

Population	*Number of missing data*	*Percent of missing data*
SESWBantu	0	0.00%
Turu	0	0.00%
Kikuyu	0	0.00%
Biaka	0	0.00%
Bakola	0	0.00%
Bamoun	0	0.00%
Yoruba	0	0.00%
Mandenka	0	0.00%
Druze	0	0.00%
Palestinian	0	0.00%
Bedouin	0	0.00%
Burusho	0	0.00%
Han	0	0.00%
Mozabite	1	0.10%
Balochi	1	0.10%
Sindhi	1	0.10%

216*Old World Genetic sample*, excluding *Masai*, *Sandawe*, *Burunge* and *Zime* from *OWF sample*. Its composition is, thus, *SESWBantu*, *San*, *Mbuti*, *Turu*, *Kikuyu*, *Biaka*, *Bakola*, *Bamoun*, *Yoruba*, *Mandenka*, *Mozabite*, *Druze*, *Palestinian*, *Bedouin*, *Hazara*, *Balochi*, *Pathan*, *Burusho*, *Makrani*, *Brahui*, *Kalash*, *Sindhi*, *Hezhen*, *Mongola*, *Daur*, *Orogen*, *Miaozu*, *Yizu*, *Tujia*, *Han*, *Xibo*, *Uygur*, *Dai*, *Lahu*, *She*, *Naxi*, *Tu*, *Cambodian*, *Japanese*, *Yakut*, *Papuan*, *NANMelanesian*, *FrBasque*, *French*, *Sardinian*, *NItalian*, *Tuscan*, *Orcadian*, *Russian* and *Adygei*, 50 populations. It must be highlighted that for these analyses using genetic data only, the *Papuan* sample was not deleted.

Population	*Number of missing data*	*Percent of missing data*
Japanese	1	0.10%
Yakut	1	0.10%
French	1	0.10%
Sardinian	1	0.10%
Russian	1	0.10%
Hazara	2	0.20%
Brahui	2	0.20%
FrBasque	2	0.20%
Pathan	3	0.31%
Adygei	3	0.31%
Orcadian	6	0.61%
Makrani	9	0.92%
Mongola	9	0.92%
Kalash	10	1.02%
Daur	10	1.02%
Tu	11	1.12%
Yizu	12	1.22%
Cambodian	13	1.32%
Orogen	16	1.63%
Tujia	16	1.63%
NItalian	16	1.63%
Hezhen	17	1.73%
She	17	1.73%
Miaozu	18	1.83%
Uygur	18	1.83%
Dai	18	1.83%
Xibo	19	1.93%
Naxi	23	2.34%
Lahu	36	3.66%
Papuan	36	3.66%
NANMelanesian	36	3.66%
Tuscan	39	3.97%
Mbuti	46	4.68%
San	139	14.14%

Table 14: The missing data analysis for populations.

The correlations (Pearson's *r*) between all pairs of genetic variants (482653) approximate very well a normal distribution with mean = 0.0239 and sd = 0.2254 (Figure 44). Unfortunately, given the limited set of populations available (50), many missing cases and very low ratio of cases to variables (0.05), it was impossible to perform a PCA analysis (Tabachnick & Fidell, 2001:582-652; Cavalli-Sforza, Menozzi & Piazza, 1994:39-42; Jobling, Hurles & Tyler-Smith, 2004:189), which would have provided a very adequate compression of the frequency information into a smaller number of principal components.

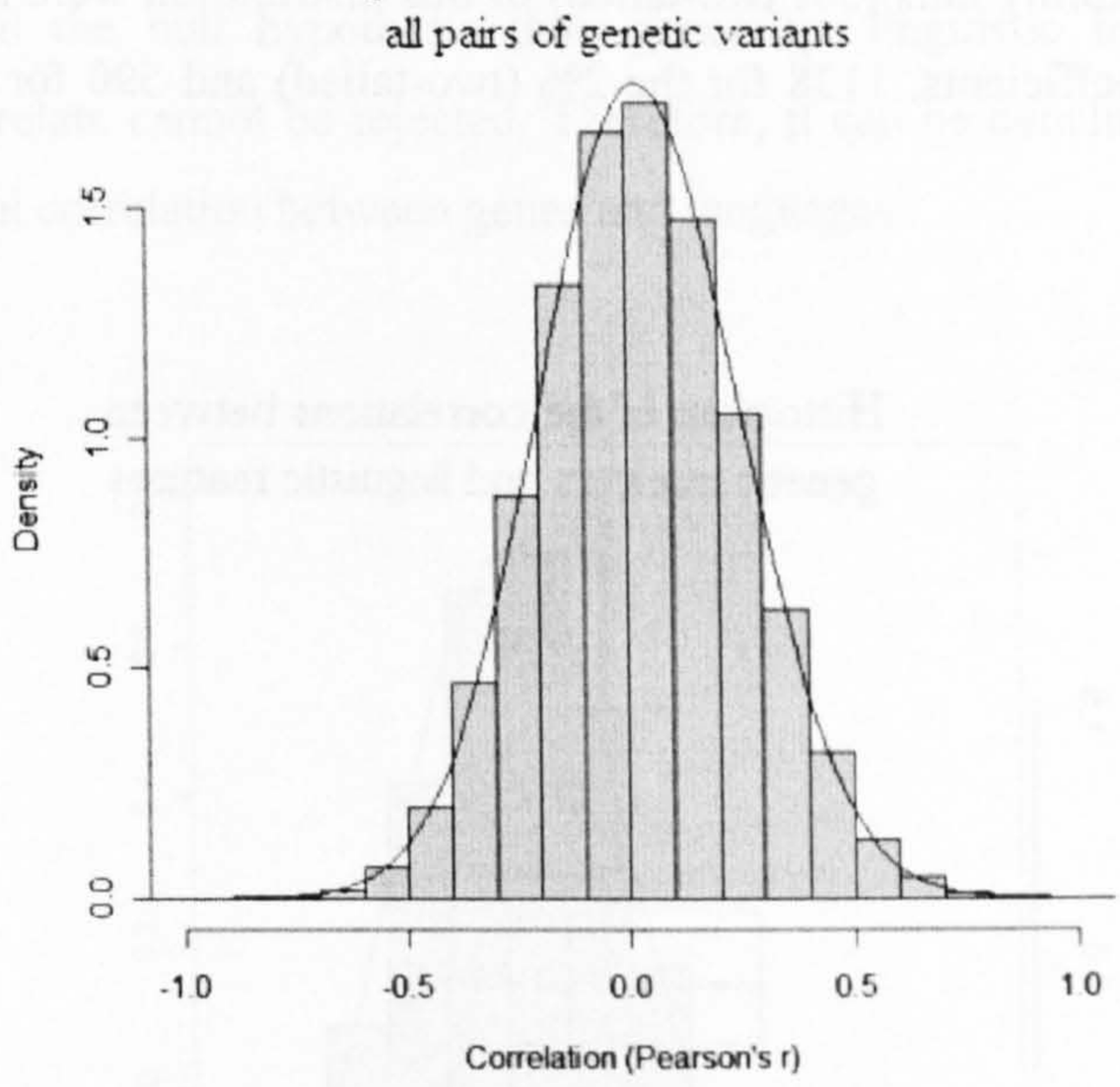

Figure 44: The correlations (Pearson's *r*) between all pairs of genetic variants.
Fits a normal distribution with mean 0.024 and sd 0.225.

4.6. Correlations between genetic variants and linguistic features

To analyze the correlations between genetic variants and linguistic features, a combined database was constructed, containing 49 populations (the *OWFinal sample*[217]), 26 linguistic features[218] and 983 genetic variants[219], denoted as the *CLGD* (Combined Linguistic-Genetic Database). For each pair (linguistic feature, genetic variant), the following measures were

217*Old World Final sample*, resulting by deleting Papuan (due to linguistic constraints, see Section 4.2) from the *OWG sample*.
218Only the binary linguistic features, Section 4.2.
219Section 4.3.

computed:

i. the Pearson's r[220] between the genetic variant's frequencies and the linguistic feature's values, and

ii. the two-samples *t-test* between the genetic variant's frequency in the two sub-populations having values 0 and 1 for the linguistic feature.

The correlation between these measures is very high (Pearson's $r = -0.9858$, $p < 2.2 \cdot 10^{-16}$), so that only the correlation coefficients (i.)[221] will be used. They fit very well a normal distribution with mean = -0.0064 and sd = 0.2183 (Figure 45), and only those correlations in the two 2.5% probability tails (5% two-tailed) of this distribution were kept. There are 2740 such correlation coefficients, 1138 for the 2% (two-tailed) and 590 for the 1% (two-tailed) tails.

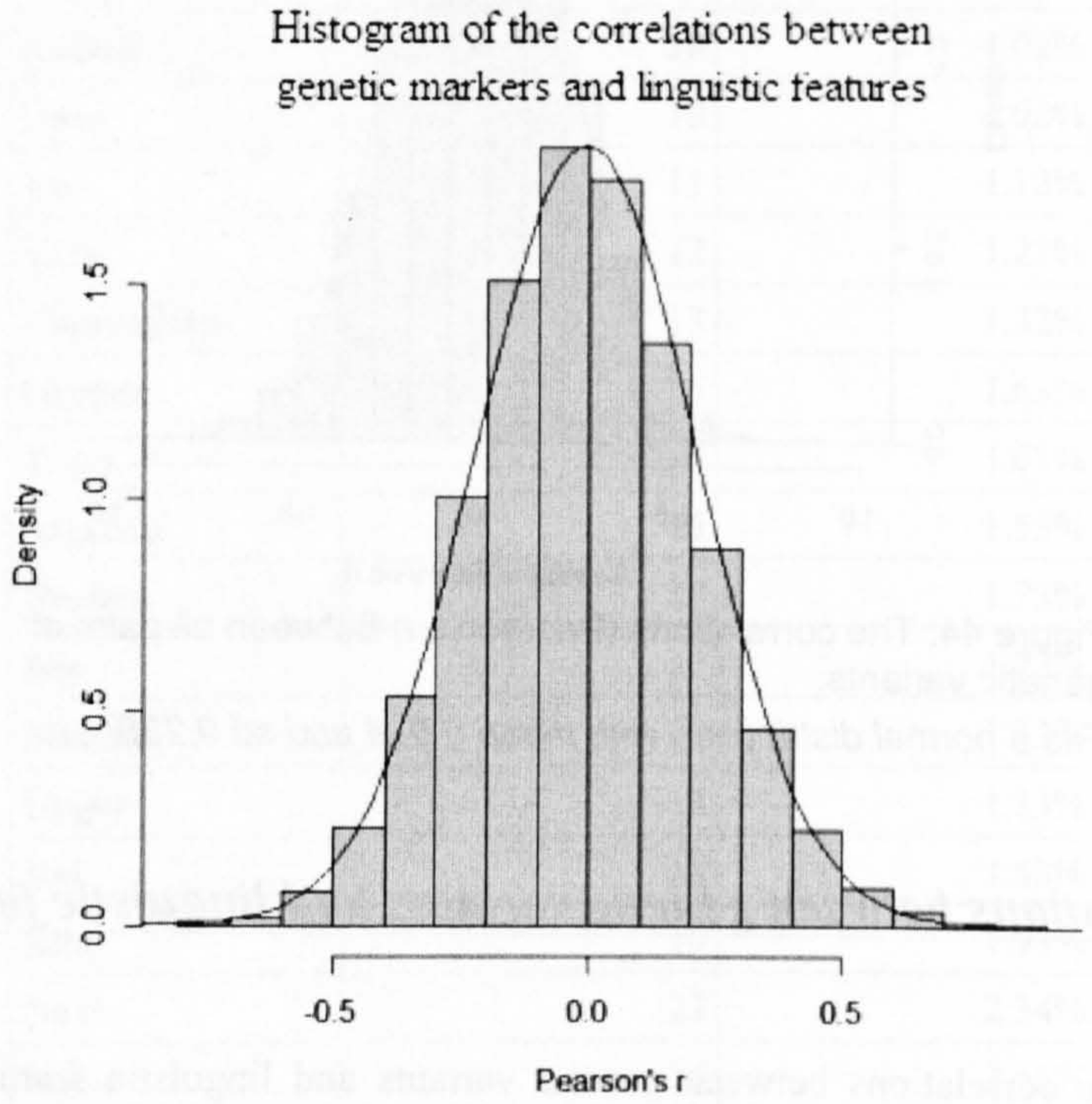

Figure 45: The correlation coefficients between genetic variants and linguistic features.

They approximate a normal curve with mean = -0.006 and sd = 0.218.

220Equivalent to the *point-biserial correlation coefficient*.

221Due to the comparability of correlation coefficients when the number of missing data varies, as opposed to the t-test values.

A valid critique[222] of this approach could be that the sign of the correlation between a binary and a continuous variable depends on the coding of the binary variable, which is non-principled in our case. Thus, the absolute values of all the Pearson's r correlation coefficients between genetic markers and linguistic features were considered and their ranks relative to the overall sample computed. The correlation between this rank and the previously fitted normal distribution is very high and significant (Pearson's $r = -0.9999$, $p < 2.2 \bullet 10^{-16}$), justifying the usage of the normally distributed correlations between genetic variants and linguistic features[223]. The distribution of the absolute values of these correlations is represented in Figure 46 and it can be seen that most of them are low (median = 0.145) and non-significant, and the null hypothesis that, generally, linguistic features and genetic variants do not correlate cannot be rejected. Therefore, it can be concluded that there is no support for a general correlation between genes and languages.

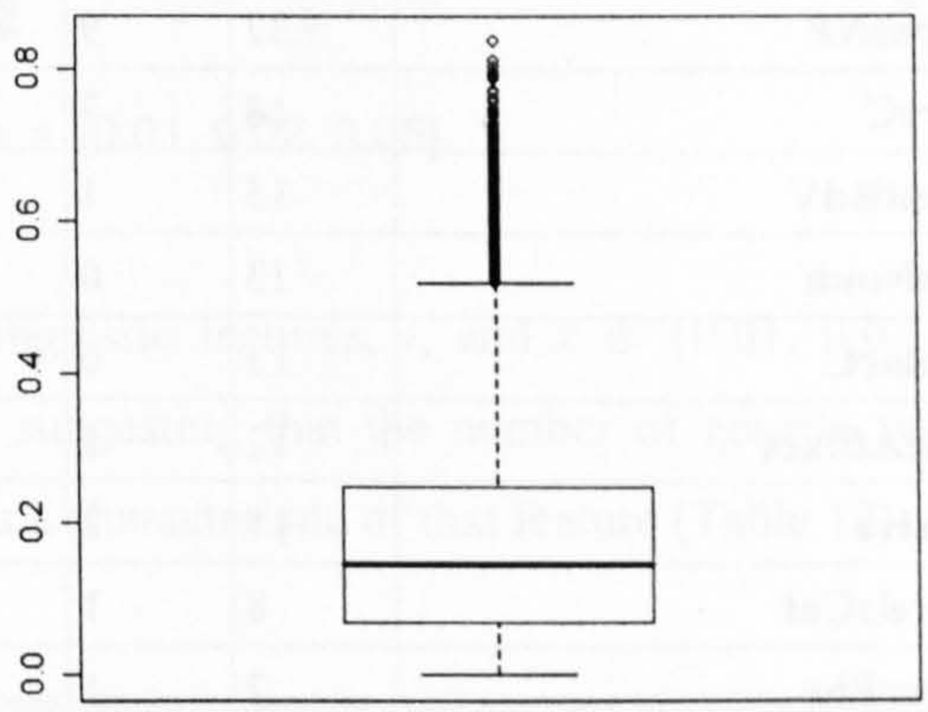

Figure 46: The boxplot of the absolute values of the correlations between genetic variants and linguistic features.

Mean = 0.1737 and median = 0.1454.

The list of genetic variants whose correlation with a given linguistic feature is in the x top (two-tailed) *of the sample of all correlations*, will be denoted[224] as $SGM_x(l)$, with $x \in \{0.01, 0.02, 0.05\}$and the[225] $|SGM_x(l)|$, for all the 26 linguistic features, l, is given in Table 15 and Figure 47.

222Thanks to Simon Kirby for pointing this out.

223And also showing that the coding of the binary linguistic features is not biased.

224Standing for *Significant genetic variants*. For example, $SGM_x(l)$ for $x = 0.05$ represents the set of all genetic variants whose correlation with linguistic feature l are in the top 5% of the empirical distribution.

225If X is a set, then $|X|$ represents the number of elements in X (the *cardinality* of X).

Linguistic feature (l) / *Level (x, two-tailed)*	*0.05*	*0.02*	*0.01*
Codas	213	119	74
NumClassifiers	200	126	81
NumNoun	152	68	43
WALSSylStr	148	78	47
Tone	119	60	27
Affixation	89	28	13
VelarNasal	78	21	2
OVWO	58	11	4
TenseAspect	56	16	6
OnsetClust	52	13	0
MorphImpv	46	16	6
AdjNoun	39	5	1
NomLoc	38	8	3
AdposNP	37	9	2
RareC	18	3	0
FrontRdV	15	1	0
GenNoun	13	0	0
UvularC	13	0	0
CaseAffixes	12	1	0
Passive	11	2	1
VowelsCat	8	1	0
InterrPhr	7	1	0
SVWO	3	2	1
ZeroCopula	2	0	0
GlotC	1	0	0
ConsCat	0	0	0
Total:	**1428**	**589**	**311**

Table 15: $|SGM_x(l)|$, for $x \in \{0.01, 0.02, 0.05\}$.

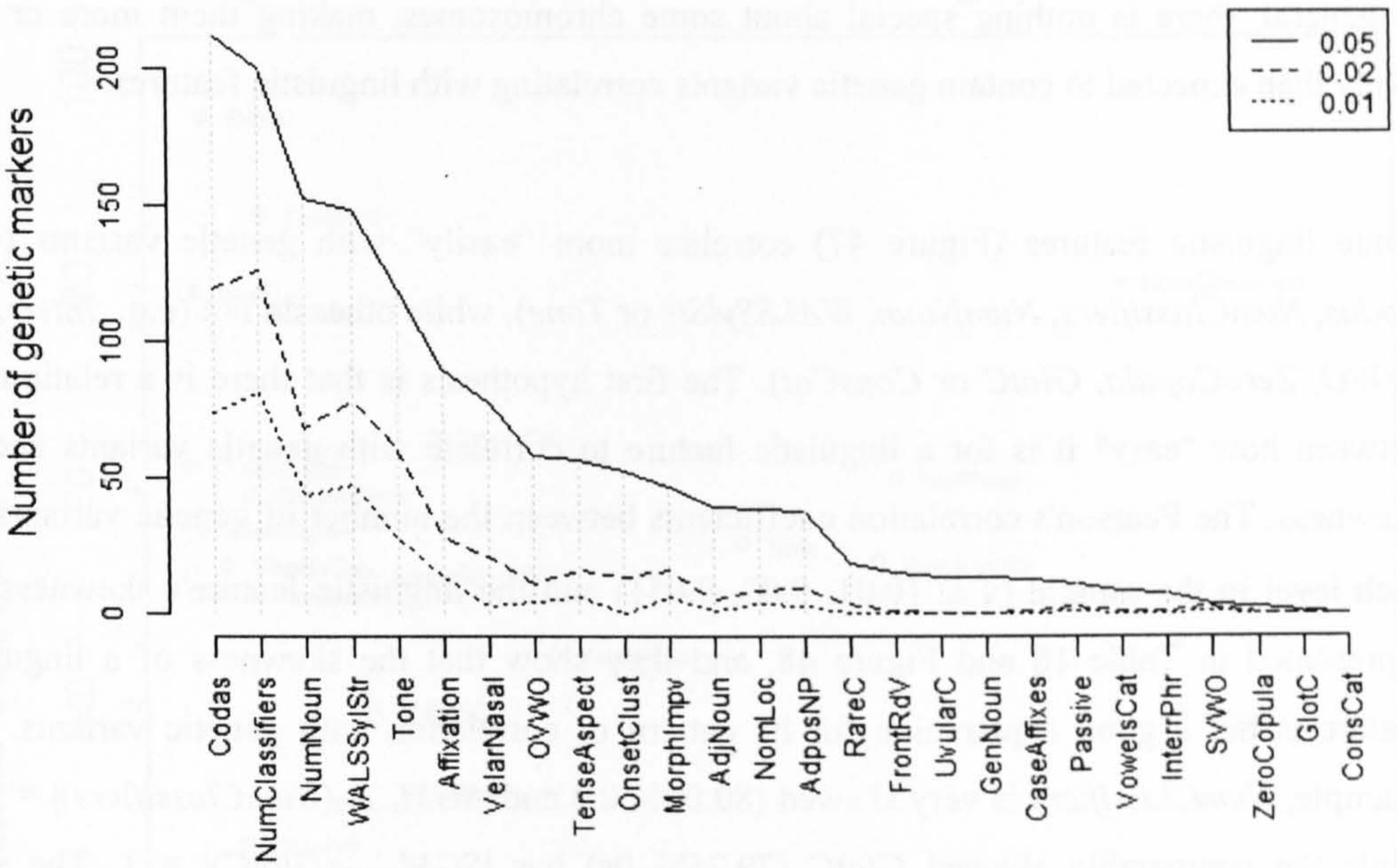

Figure 47: $|SGM_x(l)|$, for $x \in \{0.01, 0.02, 0.05\}$.

The $|SGM_x(l)|$, for all linguistic features, l, and $x \in \{0.01, 0.02, 0.05\}$, show very high correlations (Table 16), suggesting that the number of genetic variants correlating with a given linguistic feature is a characteristic of that feature (Table 17).

x	*0.02*	*0.01*
0.05	0.973	0.943
0.02		0.990

Table 16: Correlations (Pearson's *r*) between the number of markers, $|SGM_x(l)|$, for various levels, *x*, across linguistic features, *l*.

All coefficients are significant at $p < 0.01$ level (Holm mcc).

x	*Min*	*Max*	*Mean*	*SD*
0.05	0	213	54.92	62.30
0.02	0	126	22.65	36.45
0.01	0	81	11.96	23.14

Table 17: Min, max, mean and sd of $|SGM_x(l)|$ function of the level, *x*.

The χ^2 goodness-of-fit test between $|SGM_x(l)|$ and the original distribution of genetic variants

across chromosomes rejects the null hypothesis of identity of these distributions at the $p < 0.05$ level (Holm mcc) only for *AdjNoun* ($x = 0.01$) and *CaseAffixes* ($x = 0.02$), showing that, in general, there is nothing special about some chromosomes, making them more or less likely than expected to contain genetic variants correlating with linguistic features.

Some linguistic features (Figure 47) correlate more "easily" with genetic variants (e.g., *Codas*, *NumClassifiers*, *NumNoun*, *WALSSylStr* or *Tone*), while other do not (e.g., *InterrPhr*, *SVWO*, *ZeroCopula*, *GlotC* or *ConsCat*). The first hypothesis is that there is a relationship between how "easy" it is for a linguistic feature to correlate with genetic variants and its skewness. The Pearson's correlation coefficients between the number of genetic variants for each level in the sample ($x \in \{0.01, 0.02, 0.05\}$) and the linguistic feature's skewness are represented in Table 18 and Figure 48, and they show that the skewness of a linguistic feature is not a good explanation for its pattern of correlation with genetic variants. For example, *NumClassifiers* is very skewed (80.00% 0s) and $|SGM_{x=0.05}(NumClassifiers)| = 200$, while the comparably skewed *GlotC* (79.25% 0s) has $|SGM_{x=0.05}(GlotC)| = 1$. The very equilibrated *WALSSylStr* (52.94% 0s) has $|SGM_{x=0.05}(WALSSylStr)| = 148$, while the comparably equilibrated *VowelsCat* (52.83% 0s) has $|SGM_{x=0.05}(VowelsCat)| = 8$. Moreover, the two samples t-test between the skewed and equilibrated linguistic features at $x = 0.05$ cannot reject the null hypothesis that they come from the same distribution: $t = -0.0158$, $df = 9.375$, $p = 0.9877$.

x	***0.05***	***0.02***	***0.01***
Correlation with skewness	-0.152, $p = 0.459$	-0.030, $p = 0.883$	0.0142, $p = 0.945$

Table 18: The correlations between the number of genetic variants at various levels in the sample, *x*, (two-tailed) and the linguistic feature's skewness.

None is significant at the 0.05 level (Holm mcc).

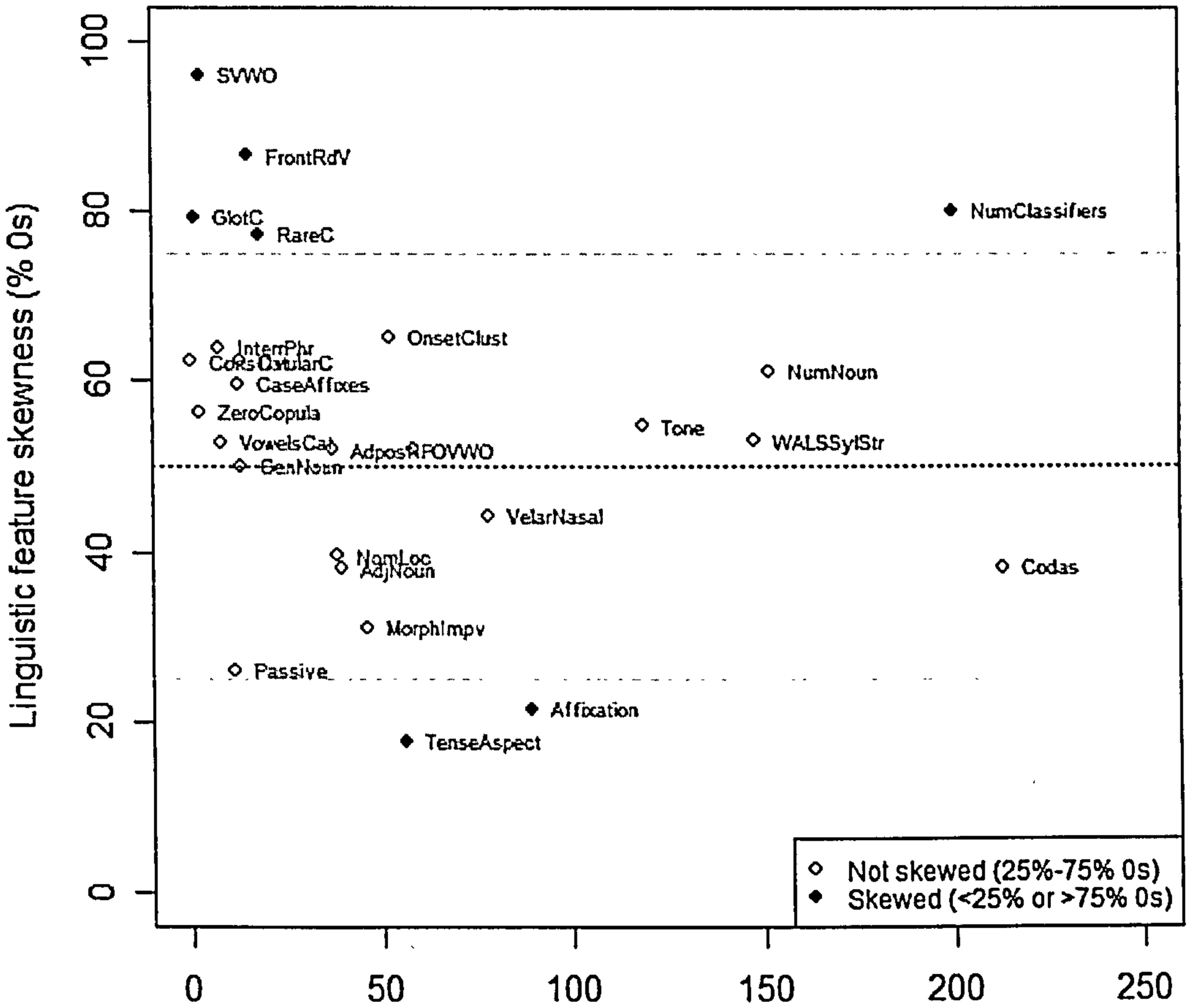

Figure 48: Scatter plot of the linguistic features' skewness versus $|SGM_{x=0.05}(l)|$.
The plots for x = 0.02 and 0.01 show the same pattern.

Table 19 presents the clustering of linguistic features based on their mutual correlations, evidencing three groups of features: a first group, composed of features correlating with many markers (213 to 119), a second group, features correlating with some genetic variants (89 to 37) and a third group, features correlating with few genetic variants (18 to 0). The first group of 5 linguistic features (*Codas*, *NumClassifiers*, *NumNoun*, *WALSSylStr* and *Tone*) falls into a tightly correlating subgroup (*Codas*, *NumNoun*, *WALSSylStr* and *Tone*) and the isolated *NumClassifiers*. (Shared genetic variants are listed in Table 20).

0.05 rank	0.02 rank	Linguistic feature	Codas (213)	NumClassif (200)	NumNoun (152)	WALSSylStr (148)	Tone (119)	Affixation (89)	VelarNasal (78)	OVWO (58)	TenseAspect (56)	OnsetClust (52)	MorphImpv (46)	AdjNoun (39)	NomLoc (38)	AdposNP (37)	RareC (18)	FrontRdV (15)	UvularC (13)	GenNoun (13)	CaseAffixes (12)	Passive (11)	VowelsCat (8)	InterrPhr (7)	SVWO (3)	ZeroCopula (2)	GlotC (1)	ConsCat (0)
1	2	**Codas (213)**			-.65	.78	-.72																					
2	1	**NumClassifiers(200)**						-.80			-.69		-.59															
3	4	**NumNoun (152)**				-.66	.73							.59													.53	
4	3	**WALSSylStr (148)**					-.73					.65																
5	5	**Tone (119)**										-.58																
6	6	**Affixation (89)**									.81		.68															
7	7	**VelarNasal (78)**																										
8	**11**	**OVWO (58)**												.56		.96				.63								
9	8	**TenseAspect (56)**											.74															
10	10	**OnsetClust (52)**																										
11	9	**MorphImpv (46)**																										
12	14	**AdjNoun (39)**														.56				.68	-.62							
13	13	**NomLoc (38)**																										
14	12	**AdposNP (37)**																		.66								
15	15	**RareC (18)**																										
16	18	**FrontRdV (15)**																										
17	**22**	**UvularC (13)**																										
18	**23**	**GenNoun (13)**																										
19	19	**CaseAffixes (12)**																										
20	**16**	**Passive (11)**																										

0.05 rank	0.02 rank	Linguistic feature	Codas (213)	NumClassif (200)	NumNoun (152)	WALSSylStr (148)	Tone (119)	Affixation (89)	VelarNasal (78)	OVWO (58)	TenseAspect (56)	OnsetClust (52)	MorphImpv (46)	AdjNoun (39)	NomLoc (38)	AdposNP (37)	RareC (18)	FrontRdV (15)	UvularC (13)	GenNoun (13)	CaseAffixes (12)	Passive (11)	VowelsCat (8)	InterrPhr (7)	SVWO (3)	ZeroCopula (2)	GlotC (1)	ConsCat (0)
21	20	**VowelsCat (8)**																										
22	21	**InterrPhr (7)**																										
23	**17**	**SVWO (3)**																										
24	24	**ZeroCopula (2)**																										
25	25	**GlotC (1)**																										
26	26	**ConsCat (0)**																										

Table 19: The significant (at the 0.05 level, Holm mcc) correlations (Pearson's *r*) between the linguistic features.

The linguistic features are ordered decreasingly by $|SGM_{x=0.05}(l)|$ (in parentheses). Column/row colors: dark gray = the 1st group of features, gray = the 2nd group and white = the 3rd group. First two columns: rank of the linguistic features for x = 0.05 and x = 0.02 levels (two-tailed) in the sample: ***bold*** *ranks on light gray background: an important (more than three places) change in rank.*

Linguistic features set	*Number of shared markers*	*Is ASPM shared?*	*Is MCPH shared?*
Codas, NumClassifiers, NumNoun, WALSSylStr, Tone	7	No	No
Codas, NumNoun, WALSSylStr, Tone **(the tightly correlating subgroup)**	58	**Yes**	No
Codas, Tone	91	**Yes**	**Yes**

Table 20: Genetic variants shared between the members of the tightly correlating groups of linguistic features, for *x* = 0.05, *x* = 0.02 and *x* = 0.01.

Due to the limited number of populations and missing data, a PCA could be performed only on the 33 genetic variants shared by *Codas*, *NumNoun*, *WALSSylStr* and *Tone* at the $x = 0.03$ level (two-tailed) in the sample and a single PC, explaining most of the variation (63%), was found (all the others are negligible) (Figure 49). The population scores on PC1 are displayed in Figure 50, and it seems to distinguish primarily between Africa and Europe, with Asia occupying an intermediate position.

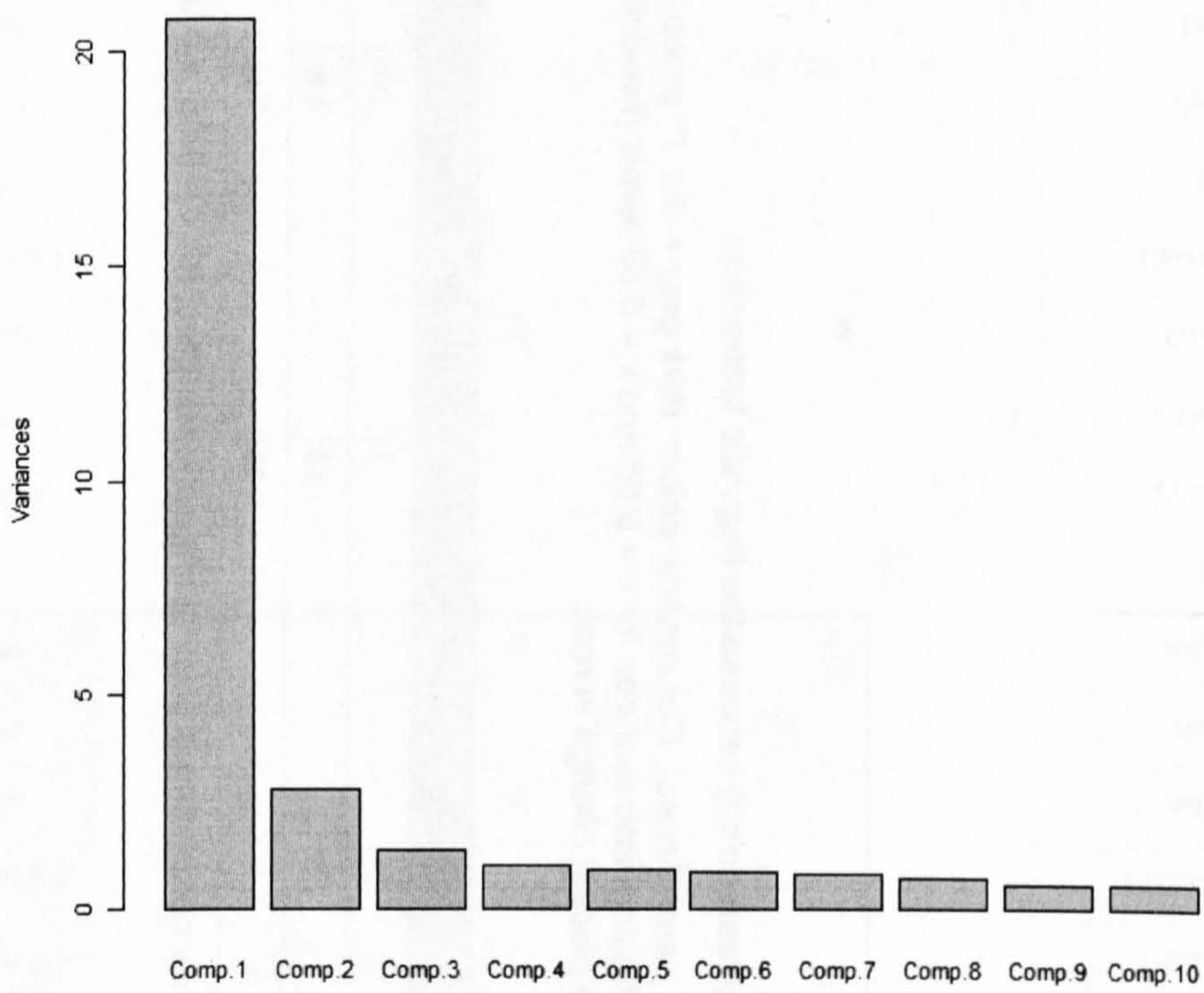

Figure 49: PCA of the 33 genetic variants shared by *Codas*, *NumNoun*, *WALSSylStr* and *Tone*.

For the x = 0.03 level (two-tailed) in the sample (only the first 10 shown). PC1 explains 63% of the variance).

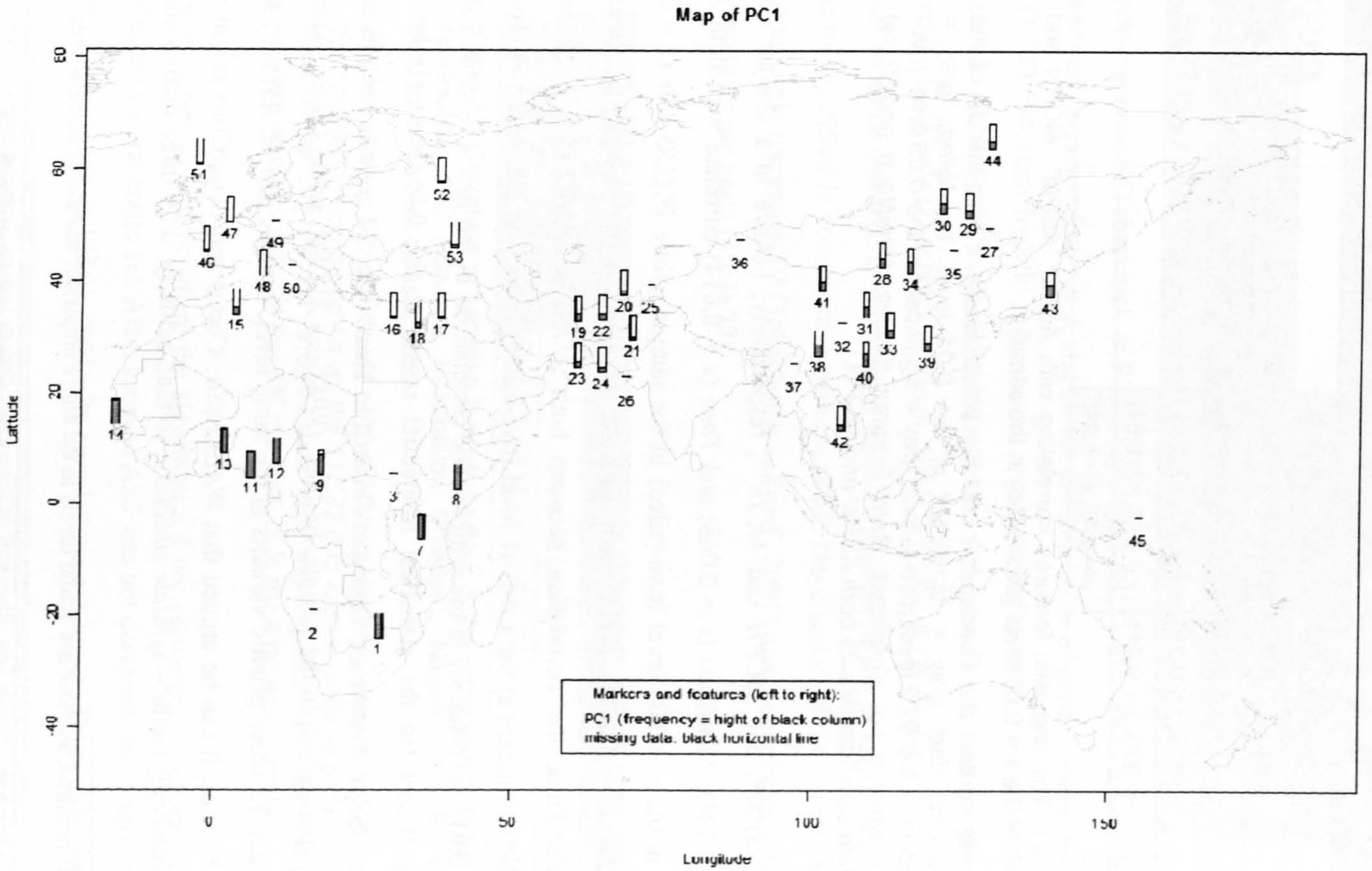

Figure 50: Map of PC1 of the 33 shared genetic variants at the $x = 0.03$ level in the sample.
It distinguishes between Europe and Africa as extremes and Asia as intermediate.

ASPM and *MCPH* are arguably the only genetic variants in the database involved in brain regulation and under positive natural selection (Evans *et al.*, 2005; Mekel-Bobrov *et al.*, 2005; Mekel-Bobrov *et al.*, 2006). The linguistic features correlating with each of them at the $x = 0.05$ level (two-tailed) in the sample, and the actual level of each such correlation[226], are listed in Table 21 below:

Linguistic feature	***Actual level (two-tailed) in the sample***			
	ASPM	***ASPM****	***MCPH***	***MCPH****
WALSSylStr	$r = 0.66, x = 0.002$	$r = 0.66, x = 0.002$		
Codas	$r = 0.63, x = 0.004$	$r = 0.63, x = 0.004$	$r = 0.68, x = 0.002$	$r = 0.68, x = 0.002$
NumNoun	$r = -0.57, x = 0.008$	$r = -0.57, x = 0.008$	$r = -0.56, x = 0.009$	$r = -0.56, x = 0.009$
Tone	$r = -0.53, x = 0.014$	$r = -0.53, x = 0.014$	$r = -0.54, x = 0.013$	$r = -0.54, x = 0.013$
OnsetClust	$r = 0.44, x = 0.041$	$r = 0.44, x = 0.041$		
NomLoc		$r = 0.43, x = 0.046$		

Table 21: The linguistic features correlating with *ASPM*, *ASPM**, *MCPH* and *MCPH** at the $x = 0.05$ level (two-tailed) in the sample.

The cells contain the Pearson's r and the actual levels, x (two-tailed), of the correlations: dark gray = significant at the 0.01 level (two-tailed), gray = significant at the 0.02 level (two-tailed), gray = significant at the 0.05 level (two-tailed), white = not significant. ***Bold*** *features (Codas, NumNoun and Tone) correlate significantly with both ASPM and MCPH.*

ASPM and *ASPM**, and *MCPH* and *MCHP**, respectively, behave in a similar manner. *Codas* ($x = 0.004$), *NumNoun* ($x = 0.008$) and *Tone* ($x = 0.014$) correlate with both *ASPM* and *MCPH* at the $x = 0.05$ level (two-tailed) in the sample, while *WALSSylStr* ($x = 0.002$) and *OnsetClust* ($x = 0.041$) correlate only with *ASPM*. *NomLoc* ($x = 0.046$) correlates only with *ASPM**. Thus, the correlation between both *ASPM* and *MCPH* and *Codas* and *NumNoun* are significant at the $x = 0.01$ level (two-tailed) in the sample, while for *Tone* the level is $x = 0.015$ (two-tailed) in the sample. The correlations in the first 4 rows of Table 21 are also significant (in the statistical inferential sense) at the 0.05 level (Holm mcc). *WALSSylStr*, *Codas*, *NumNoun*, *Tone* and *OnsetClust* share other 22 genetic variants at the $x = 0.05$ level (two-tailed) in the sample, besides *ASPM* and *ASPM**, while *Codas*, *NumNoun* and *Tone* share 72 other genetic variants at the same level, besides *ASPM*, *ASPM**, *MCPH* and *MCPH**. Thus, it can be argued that *WALSSylStr*, *Codas* and *OnsetClust* might reflect the same underlying factor – *syllable structure*, so that, actually, only three factors correlate

226No mcc is required, as these are actual levels in the entire empirical sample of correlations.

significantly with *ASPM* and *MCPH*: *Tone*, *syllable structure* and *NumNoun*. Moreover, *Tone* and *syllable structure* probably correlate on purely linguistic grounds.

Therefore, the hypothesis that there is a non-null relationship between *ASPM*, *MCPH* and *Tone* cannot be rejected at the inferential significance level $p = 0.05$, and this correlation is also in the top 5% of the empirical distribution.

4.6.1. Correlations between linguistic features and pairs of genetic variants

Nevertheless, as our hypothesis concerns the relationship between *ASPM*, *MCPH* and *Tone*, the next step is to consider pairs of genetic variants and single linguistic features. Thus, *logistic regression* (Tabachnick & Fidell, 2001:517-581) was used to assess the relationship between all linguistic features, l, and all the possible pairs of genetic variants, (g_1, g_2): There are 11,582,690 such logistic regressions, and three indicators of the *goodness of fit* were computed: *AIC* (Akaike's Information Criterion), *Nagelkerke's* R^2 and *the percent of cases correctly classified* (Tabachnick & Fidell, 2001:517-581; R Development Core Team, 2006). It must be noted that AIC is most useful when comparing nested models, the percent of cases correctly classified is not sensitive enough, while Nagelkerke's R^2 is not entirely equivalent to the multiple regression R^2 (Tabachnick & Fidell, 2001), but, nevertheless, the correlations between them are good (Table 22):

Goodness of fit measure	*Nagelkerke's R^2*	*% correct classification*
AIC	**-.566** (-0.840)	**-.557** (-0.751)
Nagelkerke's R^2		**.825** (0.771)

Table 22: Correlations between three indicators of the goodness of fit for logistic regression: *AIC* (Akaike's Information Criterion), Nagelkerke's R^2 and the percent of correct classification.

*All Pearson's r, are significant at $p < 0.01$ (two-tailed, Holm mcc). In **bold** are the correlation computed for all the 11,582,690 cases, while in parentheses the correlations only for the 87,024 cases involving linguistic features and genetic variants correlating in the top 5% (two-tailed) in the empirical sample.*

There are cases of bad fit where the algorithm did not converge, which explains the lower correlations for the entire set. Also, because AIC was not designed for comparing non-nested models (Tabachnick & Fidell, 2001), its correlations are lower. Thus, only Nagelkerke's R^2

will be used as the sole indicator of the goodness of fit for the logistic regressions. Its distribution for the entire set is represented in Figure 51 while for the 87,024 "best" cases (involving only linguistic features and genetic variants correlating in the top 5% (two-tailed) in the empirical sample) in Figure 52: these distributions are not normal, with means of 0.1457 and 0.5574 and sd's of 0.1503 and 0.1252, respectively.

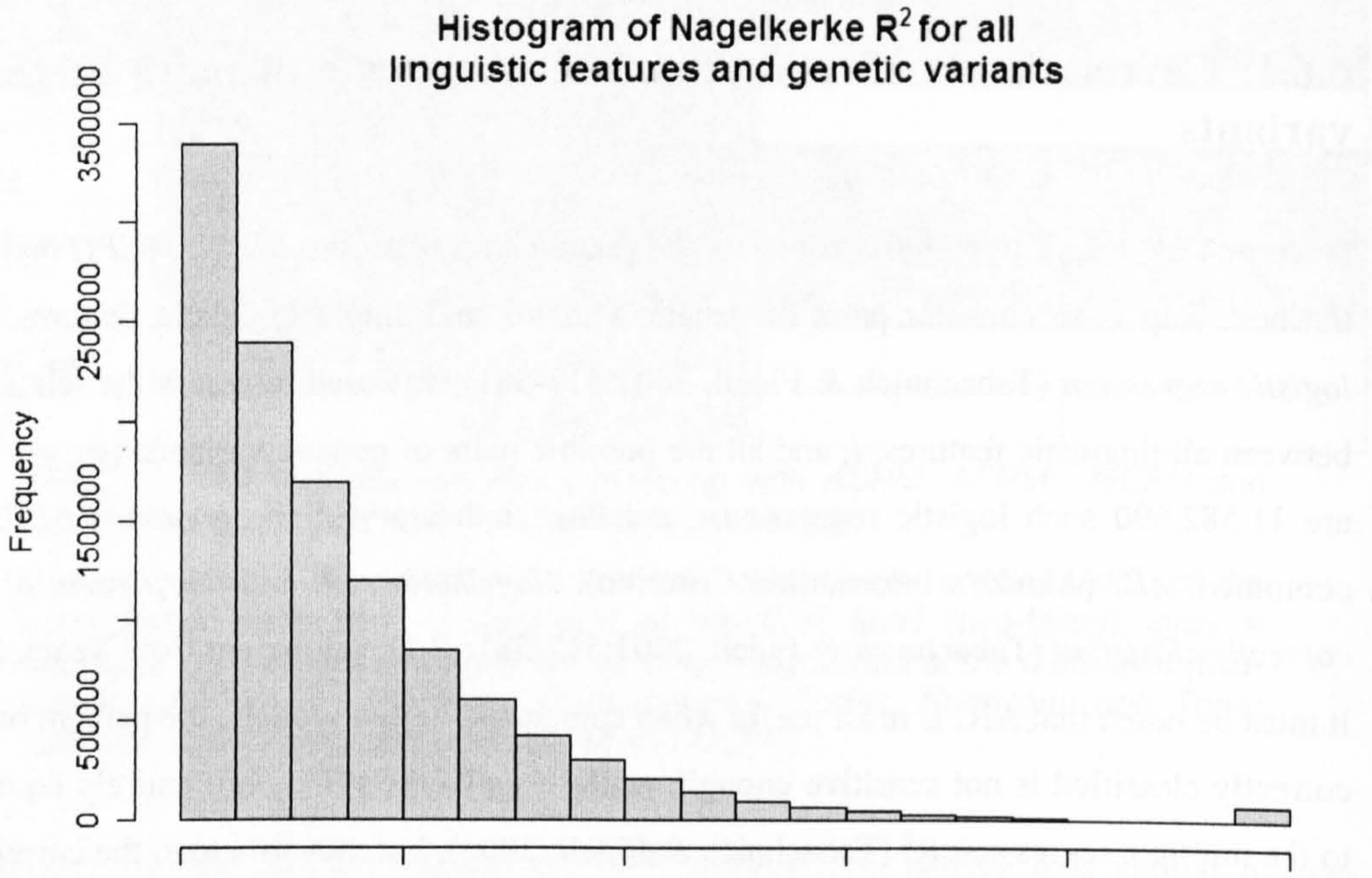

Figure 51: Histogram of the Nagelkerke's R^2 of all the 11,582,690 logistic regressions of all linguistic features on all pairs of genetic variants.

The distribution is not normal, mean=0.1456596, sd=0.1503422. The increase in frequency at the right end is due to NumClassifiers and SVWO, which are skewed and have many missing data.

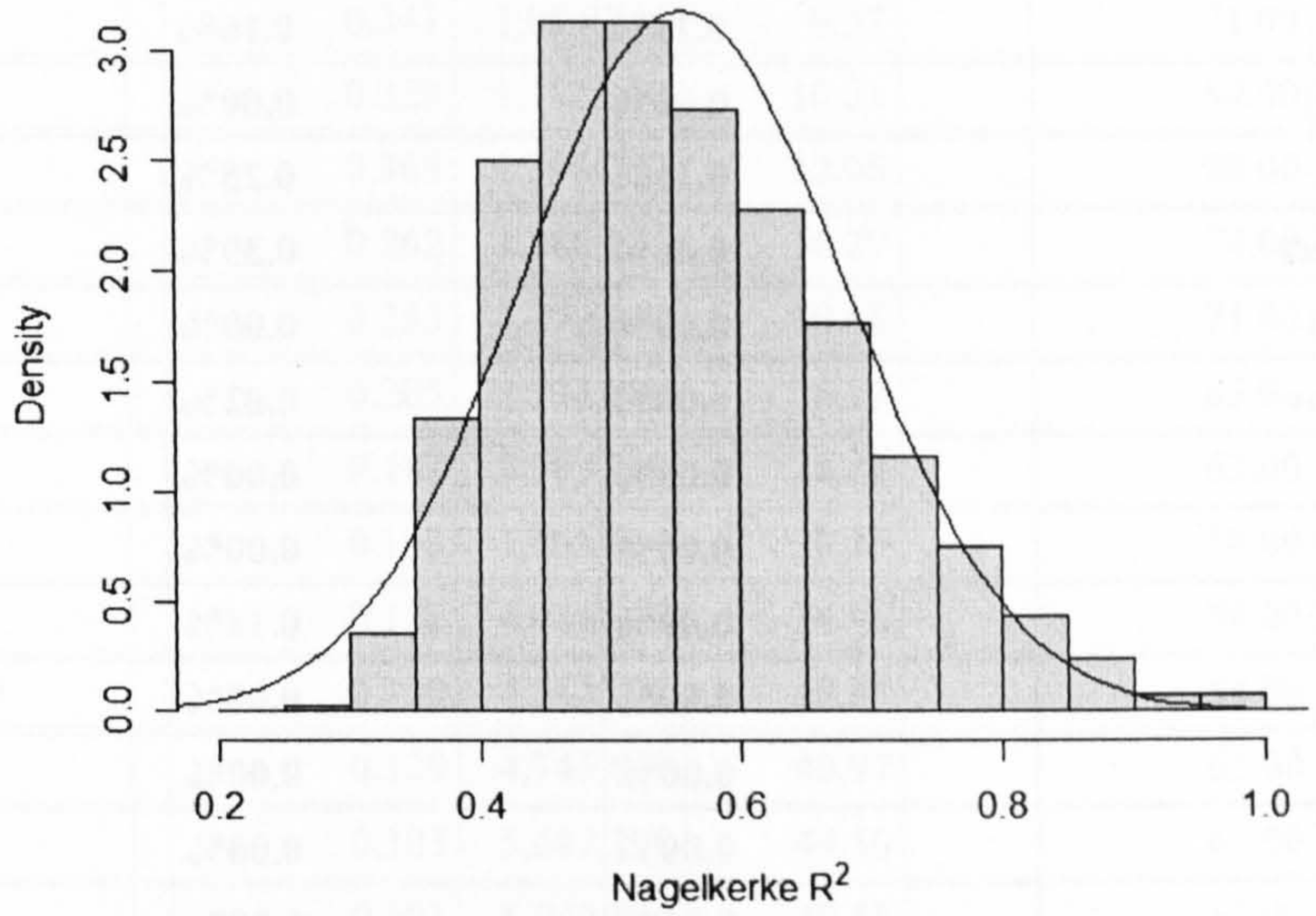

Figure 52: Histogram of the Nagelkerke's R^2 of the 87,024 "best" logistic regressions. *The distribution is not normal, mean=0.5573952, sd=0.1251867.*

The linguistic features' representation in the top 1%, 5% and all "best" logistic regressions are represented in Table 23. It can be seen that *NumClassifiers* and *SVWO* are heavily over-represented in the top 1% and 5%, while the others are under-represented, especially *Codas*, *NumNoun*, *WALSSylStr*, *Affixation* and *VelarNasal*. *Tone* is also under-represented, but, especially for the top 1%, it is reasonably close to the expectation. Over-representation in the top 1% and 5% seems to suggest that the pattern of these linguistic features is "easily" described by using just two genetic variants, while those under-represented features seem to have a pattern too complex to be captured by just two genetic variants. The linguistic features present in top 1% are: *NumClassifiers*, *SVWO*, *NumNoun*, *Codas*, *TenseAspect*, *WALSSylStr*, *Tone*, *OVWO & Affixation* and those present only in top 5% are: *AdposNP*, *MorphImpv & Passive*.

Linguistic feature	*% in top 1%*	*% in top 5%*	*% overall*
Codas	**6.78%**	**13.97%**	**25.95%**
NumClassifiers	**83.10%**	**70.97%**	**22.87%**
NumNoun	**3.10%**	**4.90%**	**13.19%**
WALSSylStr	**0.57%**	**5.91%**	**12.50%**

Linguistic feature	*% in top 1%*	*% in top 5%*	*% overall*
Tone	**5.40%**	**3.17%**	**8.07%**
Affixation	**0.11%**	**0.16%**	**4.50%**
VelarNasal	**0.00%**	**0.00%**	**3.45%**
OVWO	**0.11%**	**0.25%**	**1.90%**
TenseAspect	**0.46%**	**0.39%**	**1.77%**
OnsetClust	**0.00%**	**0.00%**	**1.52%**
MorphImpv	**0.00%**	**0.02%**	**1.19%**
AdjNoun	**0.00%**	**0.00%**	**0.85%**
NomLoc	**0.00%**	**0.00%**	**0.81%**
AdposNP	**0.00%**	**0.14%**	**0.77%**
RareC	**0.00%**	**0.00%**	**0.18%**
FrontRdV	**0.00%**	**0.00%**	**0.12%**
UvularC	**0.00%**	**0.00%**	**0.09%**
GenNoun	**0.00%**	**0.00%**	**0.09%**
CaseAffixes	**0.00%**	**0.00%**	**0.08%**
Passive	**0.00%**	**0.05%**	**0.06%**
VowelsCat	**0.00%**	**0.00%**	**0.03%**
InterrPhr	**0.00%**	**0.00%**	**0.02%**
SVWO	**0.34%**	**0.07%**	**0.00%**
ZeroCopula	**0.00%**	**0.00%**	**0.00%**
GlotC	**0.00%**	**0.00%**	**0.00%**
ConsCat	**0.00%**	**0.00%**	**0.00%**

Table 23: For each linguistic feature, its representation (percents) in the top 1% and 5% "best" logistic regressions (Nagelkerke's R^2) and overall.

The goodness of fit indicators of the logistic regressions of linguistic features on *ASPM* and *MCPH* are:

Linguistic feature	*Nagelkerke's R^2*	*Rank*	*Percent*	*% correctly classified*	*AIC*
Codas	**0.644**	**156,559**	**1.35**	**84.00**	**0.00**
WALSSylStr	**0.579**	**226,835**	**1.96**	**85.00**	**0.00**
NumNoun	**0.558**	**254,703**	**2.20**	**78.00**	**0.00**
Tone	**0.528**	**303,709**	**2.62**	**73.00**	**0.00**
VelarNasal	0.350	1,014,811	8.76	69.00	0.00
NumClassifiers	0.346	1,048,031	9.05	83.00	0.00

Linguistic feature	*Nagelkerke's R^2*	*Rank*	*Percent*	*% correctly classified*	*AIC*
FrontRdV	0.341	1,084,746	9.37	71.00	0.00
RareC	0.329	1,182,369	10.21	69.00	0.00
NomLoc	0.305	1,396,353	12.06	80.00	0.00
OnsetClust	0.262	1,886,725	16.29	74.00	0.00
OVWO	0.235	2,279,948	19.68	71.00	0.00
AdjNoun	0.205	2,794,882	24.13	65.00	0.00
UvularC	0.183	3,269,210	28.22	63.00	0.00
CaseAffixes	0.173	3,503,804	30.25	58.00	0.00
Affixation	0.152	4,051,322	34.98	58.00	0.00
TenseAspect	0.129	4,742,978	40.95	54.00	0.00
AdposNP	0.129	4,745,659	40.97	65.00	0.00
GenNoun	0.103	5,687,198	49.10	61.00	0.00
SVWO	0.101	5,742,584	49.58	63.00	0.00
MorphImpv	0.077	6,826,178	58.93	60.00	0.00
VowelsCat	0.064	7,419,075	64.05	59.00	0.00
InterrPhr	0.053	8,034,343	69.37	60.00	0.00
Passive	0.036	9,039,128	78.04	67.00	0.00
ZeroCopula	0.017	10,276,493	88.72	52.00	0.00

Table 24: The goodness of fit indicators of the logistic regressions of linguistic features on *ASPM* and *MCPH*.

The rank represents the overall rank of Nagelkerke's R^2 in the entire set of 11,582,690 cases, while the percent represents the percent of cases better than it. In gray ***bold*** *the linguistic features in the top 5% Nagelkerke's R^2.*

The correlation between the ranks of linguistic features for these and those corresponding to *ASPM** and *MCPH** (the two genetic variants derived from *ASPM* and *MCPH* applying the missing data procedure) is very high ($r = 0.9984$, $p < 2.2 \bullet 10^{-16}$), supporting, again, the adequacy of the African missing data handling procedure. The logistic regression of *Codas*, *WALSSylStr*, *NumNoun* and *Tone* on *ASPM* and *MCPH* are in the top 5% logistic regressions of all linguistic features on all pairs of genetic variants. For *Codas*, *WALSSylStr* and *Tone*, the *ASPM* and *MCPH* explain 65%, 58% and 52% of the variance, respectively, and the percents of correct classifications are 84%, 85% and 73%, suggesting that the pair *ASPM-MCPH* is a very good predictor for these linguistic features; their logistic regression coefficients[227] are listed in Table 25.

227The logistic regression equation is (Tabachnick & Fidell, 2001:523): $Y = e^X/(1+e^X)$, where $X = A +$

Linguistic feature	*Intercept (A)*	B_{ASPM}	B_{MCPH}
Codas	-3.949	7.166	4.776
WALSSylStr	-4.263	10.863	2.818
Tone	4.478	-7.170	-4.952

Table 25: The logistic regression coefficients.

It can be concluded, thus, that the hypothesis of a non-null relationship between *ASPM*, *MCPH* and *Tone* is also supported when *ASPM* and *MCPH* are treated as a pair.

4.7. Controlling for geography: spatial analyses of genetic variants and linguistic features

When relating spatially patterned distributions, it is very important to control for the influence of space itself on the relationship. More exactly, let's consider two variables with a non-random spatial patterning and a non-null correlation between them. In this case, one must also consider the partial correlation between them when controlling for space, as the spatial relationships between variables can either have no effect, add or subtract from the "true" correlation. The problematic of incorporating space in the relationship between variables is very complex and is treated in an accessible manner in Fortin & Dale (2005) and Upton & Fingleton (1985).

4.7.1. Geographic, genetic and linguistic distances

Therefore, the relationships between geography, on one hand, and genetic and linguistic distributions, on the other, was analyzed in order to understand its role in shaping these diversities. The main geographical assumption is that the routes relevant for genetic and linguistic diversities are located, as much as possible, on land. *Land distances* were approximated using *great circle distances*[228] for pairs of locations on the same continent, while forcing the intercontinental paths to pass through specific *connection points*: Damascus (33°30'N, 36°19'E) for Africa/Eurasia, Bangkok (13°45'N, 100°30'E) for

---

$B_1X_1 + B_2X_2 + ... + B_nX_n$, with Y the DV and $X_1, X_2, ..., X_n$ the IVs.

228 Defined as the shortest distance between two points on a sphere (e.g., http://mathworld.wolfram.com/GreatCircle.html, http://williams.best.vwh.net/gccalc.htm, August, 2006)

Melanesia/Eurasia and Fairbanks (64°49'N, 147°45'W) for Americas/Eurasia.

The *genetic distances* between populations were computed using Nei's (Nei, 1972; Jobling, Hurles & Tyler-Smith, 2004:168) standard genetic distance, D, defined as $D = -\ln I$, where I is the *identity of genes between the two population* (Nei, 1972:284). The actual formula used is based on Jobling, Hurles & Tyler-Smith (2004:168) and, especially, on Felsenstein (2005 – GENDIST in PHYLIP 3.6):

$$D_G = -\ln \frac{\sum p_{1mi} p_{2mi}}{\sqrt{(\sum p_{1mi}^2)(\sum p_{2mi}^2)}}$$

where m is summed over all loci, i over all the alleles at the m^{th} locus, p_{1mi} is the frequency of the i^{th} allele at the m^{th} locus in population 1 and p_{2mi} is the frequency of the i^{th} allele at the m^{th} locus in population 2.

For the *linguistic distances*, a simple, a-theoretical approach was used: for any set of linguistic features, $f_1, f_2, ..., f_n$, and pair of populations, p_1 and p_2, the linguistic distance is given by the standard Euclidean distance in an n-dimensional space:

$$D_L(f_1, f_2, ..., f_n; p_1, p_2) = \sqrt{(\sum (f_{1i} - f_{2i})^2)}$$

where i is summed over all linguistic features $1..n$, f_{1i} is the value of the i^{th} linguistic feature in population 1 and f_{2i} is the value of the i^{th} feature in population 2. To test the intuition that some linguistic features "are more predictable" than others and could thus impact more on the distance between languages, three weighting schemes were used: equal weighs, weights proportional to the informational entropy of the linguistic features and weights inversely proportional to this information entropy. Thus, the *generalized linguistic distance* becomes:

$$D_L(f_1, f_2, ..., f_n; p_1, p_2; w_1, w_2, ..., w_n) = \sqrt{(\sum w_i (f_{1i} - f_{2i})^2)}$$

The *equal weighting scheme* (EWS) simply considers all linguistic features equally important,

$$w_1 = w_2 = ... = w_n = 1/n$$

where n is the number of features. The other two weighting schemes are based on the "informational content" of a linguistic feature, as measured by its *informational entropy* (Shannon, 1948). Let v_i be the frequency of 1s for linguistic feature f_i, then its informational entropy is given by:

$$H_i = -[v_i \log v_i + (1-v_i)\log(1-v_i)]$$

v_i and H_i for the 26 linguistic features are listed in Table 26:

Linguistic feature	*Frequency of 1s (v)*	*Informational entropy (H)*
VowelsCat	0.49	0.99970
WALSSylStr	0.51	0.99967
GenNoun	0.48	0.99851
OVWO	0.46	0.99498
AdposNP	0.46	0.99498
VelarNasal	0.56	0.98870
AdjNoun	0.59	0.97807
ZeroCopula	0.41	0.97602
Tone	0.41	0.97553
CaseAffixes	0.40	0.96846
UvularC	0.39	0.96334
OnsetClust	0.38	0.96124
NomLoc	0.63	0.95443
InterrPhr	0.37	0.95227
Codas	0.63	0.94945
ConsCat	0.35	0.93130
NumNoun	0.33	0.91830
MorphImpv	0.67	0.91830
Affixation	0.77	0.77656
NumClassifiers	0.22	0.75928
Passive	0.78	0.75538
RareC	0.20	0.73002
TenseAspect	0.81	0.69621
FrontRdV	0.15	0.59931
GlotC	0.14	0.59167
SVWO	0.02	0.15110

Table 26: For each linguistic feature: the frequency of 1s and its informational entropy, *H*.

The *directly proportional to the informational entropy of the linguistic features weighting scheme* (DPWS) considers more important those features which carry more information (their distribution is less skewed). In this case, the weight of feature f_i is

$$w_i = H_i / \Sigma H_i$$

normalized so that they sum up to 1. The *inversely proportional to the informational entropy of the linguistic features weighting scheme* (IPWS) considers more important those features whose distribution is more skewed, as any two random languages are less likely to differ in the value of such a linguistic feature. In this case, the weight of feature f_i is

$$w_i = 1/(H_i \Sigma(1/H_i))$$

normalized so that they sum up to 1.

4.7.2. Correlations between distance matrices: the Mantel correlation

Given two distance matrices (or, more generally, two similarity or dissimilarity matrices), a correlation coefficient between these distances can be computed, using the approach known as *Mantel correlation* or *Mantel test* (Mantel, 1967). This consists of the computation of a classical correlation coefficient, like Pearson's *r* or Spearman's *rho*, (Howitt & Cramer, 2003:67) between the elements of the two matrices and the subsequent testing of the null hypothesis of no relationship through a randomization procedure (Edgington, 1987), whereby the rows and columns of one matrix are randomly shuffled. It must be pointed out that the *p*-values of the standard correlation coefficients are inadequate for distance matrices, because the cells are not independent. A *partial Mantel test* computes the correlation between two matrices when controlling for the effects of a third. Description of the procedure and discussions are given, for example, in Bonnet & Van de Peer (2002:2-3) and Fortin & Dale (2005:147-153). The Mantel and partial Mantel tests are very useful in, for example, assessing the association between genetic and linguistic distances when controlling for geographical distances. Nevertheless, It must be pointed out that the interpretation of Mantel and partial Mantel coefficients (r_{AB} and $r_{AB.C}$, respectively) is fraught with difficulties, as they refer not to the relationship between values but between distances between values (Fortin & Dale, 2005:149). In the following, (partial) Mantel tests were performed using the method of Legendre & Legendre (1998)[229]. Also, a visualization technique for displaying distance matrices through gray scales will be used, whereby the color of each cell ranges from black (minimum) to white (maximum)[230].

229As implemented by R's (R Development Core Team, 2006) `vegan` package, methods `mantel` and `mantel.partial` with 10,000 random permutations.

230As implemented by R's (R Development Core Team, 2006) `color2D.matplot` method of package `plotrix`.

To evaluate the differences between the three methods of computing linguistic distances (EWS, DPEW and IPWS), the Mantel correlations between these linguistic distance matrices, both for all features, and for *Tone*, *Codas* and *WALSSylStr* only (see below), were computed (Table 27).

Weighting method	***EWS***	***DPWS***	***IPWS***
EWS		0.996	0.978
DPWS	*1.000*		0.959
IPWS	*1.000*	*0.999*	

Table 27: Mantel correlations between linguistic distance matrices computed using the three weighting schemes.

Upper triangle: for all linguistic features; lower triangle, italic: Tone, Codas and WALSSylStr only. All correlations significant at p < 0.001 (Holm mcc).

The three weighting schemes are essentially equivalent, suggesting that the linguistic distance between populations is not affected disproportionately by certain features at the expense of others; thus, only the EWS will be used.

For the entire set of populations[231], genetic variants and linguistic features, the following distance matrices were obtained (Figures 53, 54 and 55):

231It would have been interesting to perform the same analyses at a *macro-regional level*, but given the small sample size available and the difficulties in delimiting such macro-areas in a way which does not inject the expected results into the assumptions, I decided to postpone them until better samples will be available.

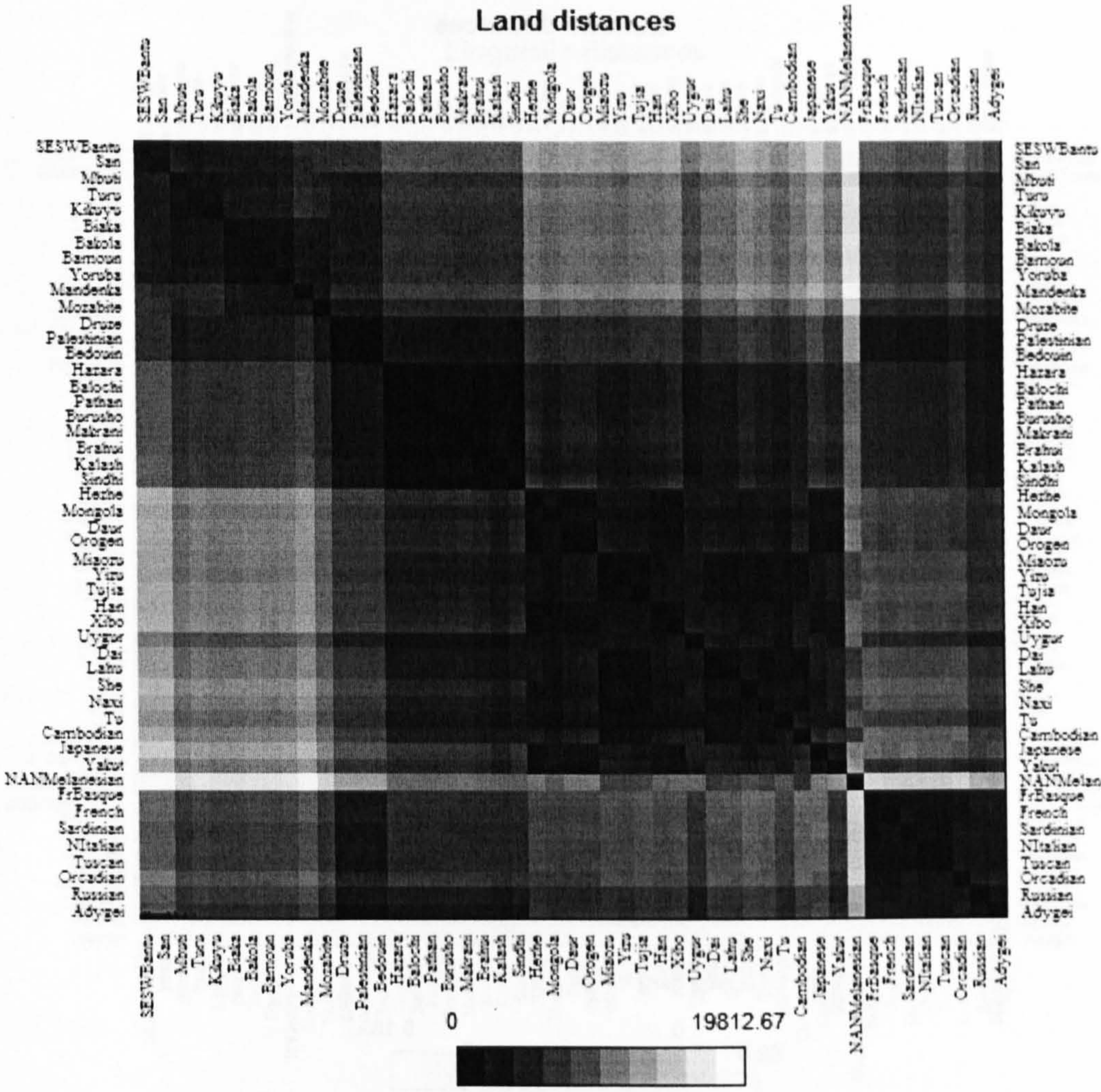

Figure 53: The land distances matrix: black (0 km) to white (19813 km).

The continental clusters are clearly visible (Africa, Europe, Asia) and the most isolated population is NANMelanesian.

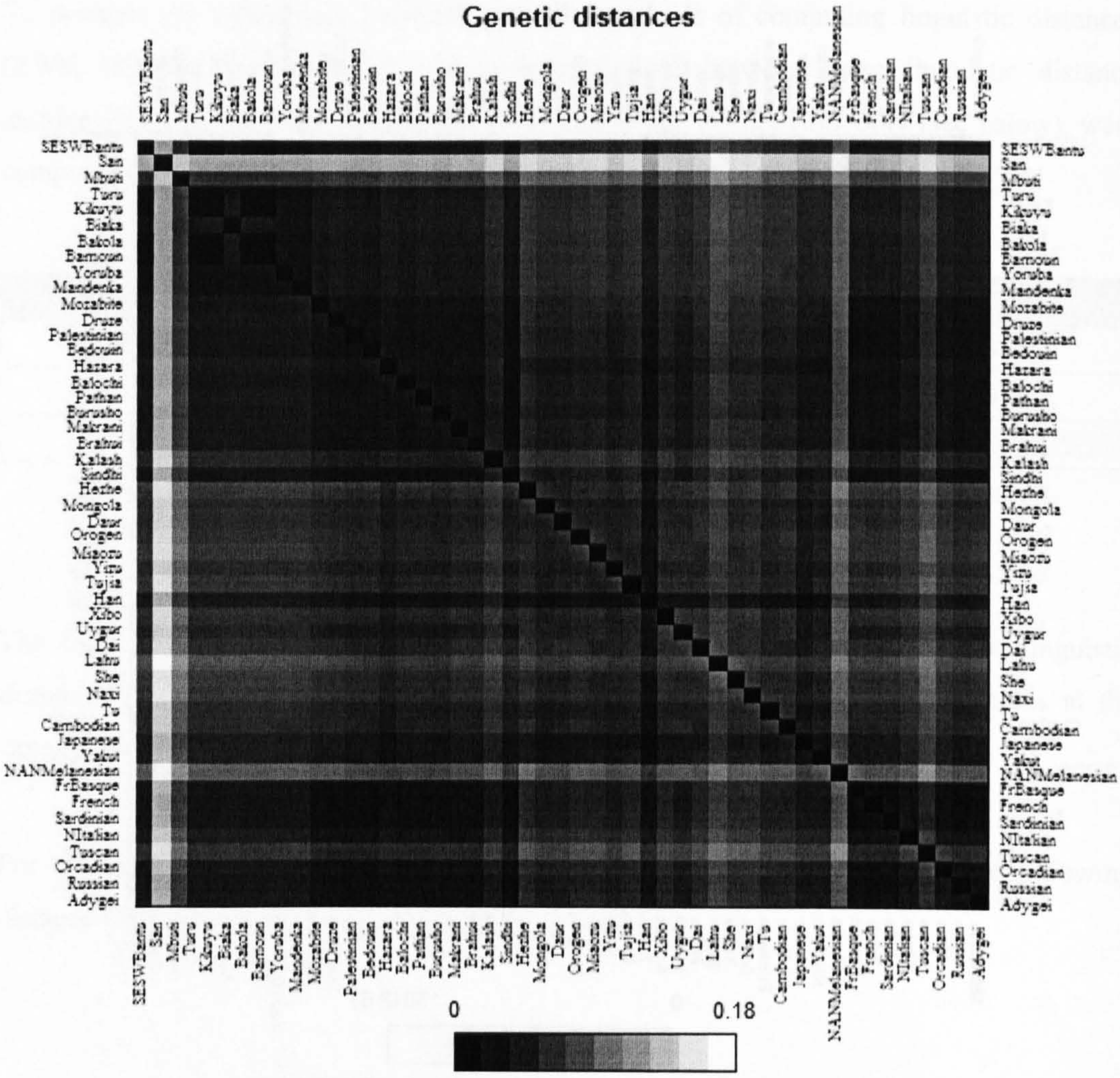

Figure 54: The genetic distances (Nei's *D*) matrix: black (0) to white (0.18).

San is a clear genetic outlier, followed by NANMelanesian. Han is an interesting case , seemingly equally distant from all other populations, but it might be due to some form of biased sampling.

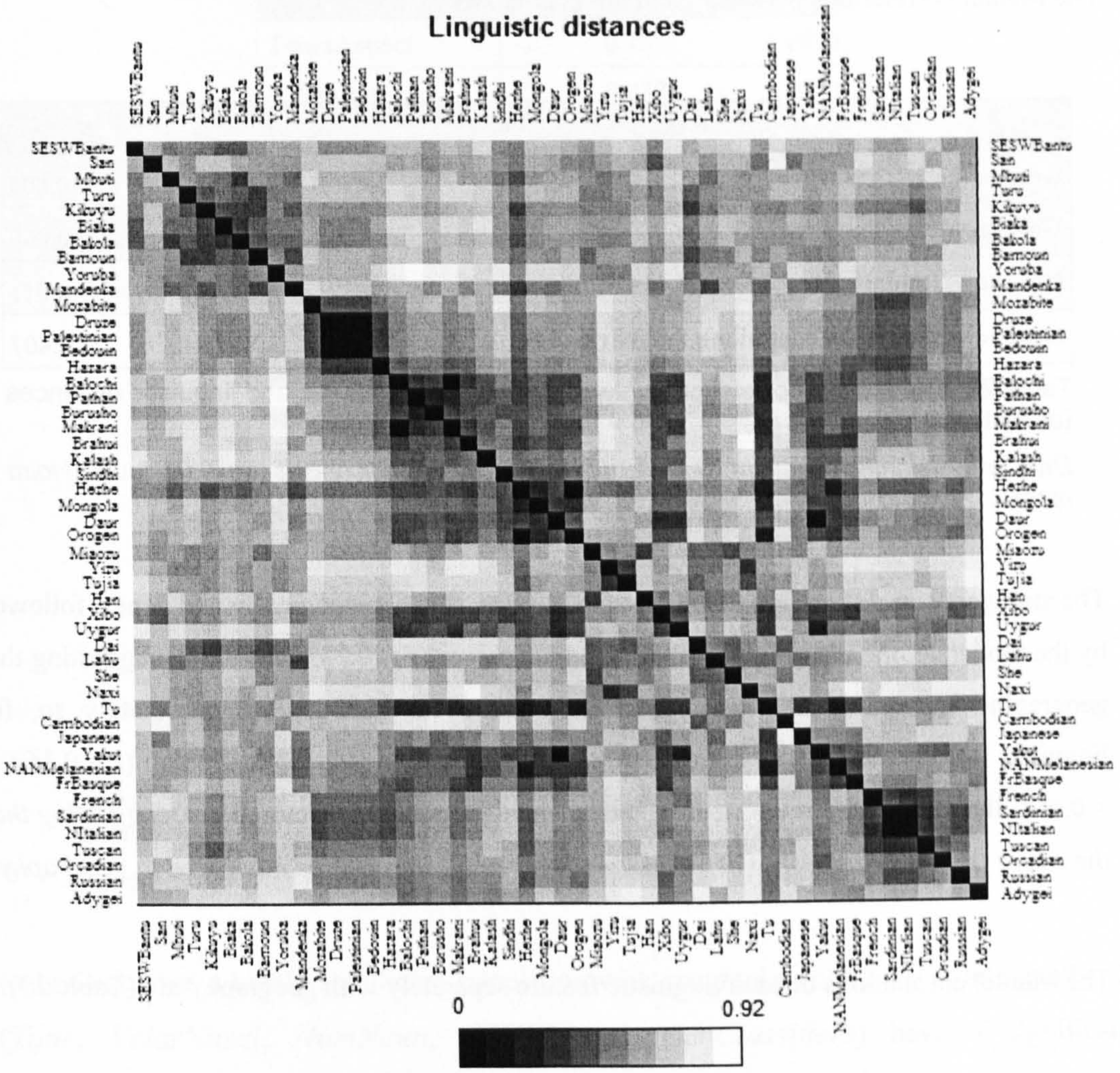

Figure 55: The linguistic distances: black (0) to white (0.92).

There seem to exist few visible patterns, except the apparent closeness of Afro-Asiatic and W. European languages. Interestingly, some S. and E. Asiatic languages seem to lay close together (Balochi to Orogen). The overall impression is of inhomogeneity as opposed to the high genetic homogeneity (see Figure 54 above).

The Mantel correlations between them are (Table 28):

Distances	*r*	*p*
Geographic vs. **genetic**	0.509	0.000
Geographic vs. **linguistic**	0.283	0.000
Genetic vs. **linguistic**[232]	0.162	0.011
Genetic vs. **linguistic** controlling for **geographic**	0.021	0.407

Table 28: The Mantel correlations between geographic, genetic and linguistic distances (all features).

Dark gray: significant at the 0.01 level, light gray: significant at the 0.05 level (Holm mcc).

The strongest correlation ($r = 0.509$, $p < 0.01$) exists between geography and genes, followed by the moderate one between geography and languages ($r = 0.283$, $p < 0.01$), suggesting that geographic separation is a very good explanation for genetic, but moderately so, for linguistic differentiation. The correlation between genes and languages is small ($r = 0.162$, $p < 0.05$), and vanishes ($r = 0.021$, $p = 0.407$) when geography is factored out, *suggesting that the entire (small) relationship between languages and genes can be attributed to geography.*

The Mantel correlations of each linguistic feature separately with geography are (Table 29):

Linguistic feature	*r*	*Adjusted p*
Codas	0.218	0.0026
Tone	0.169	0.0150
VelarNasal	0.152	0.0240
NumNoun	0.181	0.0253
OVWO	0.162	0.0264
NumClassifiers	0.209	0.0399
AdposNP	0.103	0.1260
RareC	0.184	0.1482
NomLoc	0.114	0.2484
WALSSylStr	0.045	1.0000

232This is a different aspect of language-genes correlations from the one discussed in Section 4.6. above. Here, distances between languages and genetic pools show a small correlation, while in the other case, there is no correlation, in general, between a linguistic feature and a genetic variant.

Linguistic feature	*r*	*Adjusted p*
TenseAspect	0.106	1.0000
Affixation	0.089	1.0000
Passive	0.080	1.0000
GlotC	0.097	1.0000
InterrPhr	0.048	1.0000
CaseAffixes	0.033	1.0000
AdjNoun	0.034	1.0000
MorphImpv	0.047	1.0000
SVWO	0.044	1.0000
ZeroCopula	0.026	1.0000
VowelsCat	0.003	1.0000
FrontRdV	0.004	1.0000
OnsetClust	-0.001	1.0000
GenNoun	0.001	1.0000
ConsCat	-0.014	1.0000
UvularC	-0.056	1.0000

Table 29: Mantel correlations between geography and each linguistic feature separately.

Adjusted p: p after Holm mcc. Gray: significant at the 0.05 level.

Only *Codas* has a highly significant correlation with geography ($p < 0.01$), while another 5 (*Tone*, *VelarNasal*, *NumNoun*, *OVWO* and *NumClassifiers*) have a significant correlation ($p < 0.05$). For these 6, the minimum is 0.152 (*VelarNasal*) and the maximum is 0.218 (*Codas*), mean = 0.182. These suggest that *some linguistic features are more strongly influenced by geographic distance than others* and, given that all significant such correlations are positive, in general, *linguistic similarity decreases with increasing spatial separation*.

After Holm mcc, 114 of the Mantel correlations between genetic variants with geography remain significant at the $p < 0.05$ level (minimum = 0.2451, maximum = 0.6264, and mean = 0.4009). *ASPM* ($r = 0.074$, *adjusted* $p = 1.0$) and *ASPM** ($r = 0.071$, *adjusted* $p = 1.0$) have non-significant correlations with geography, while *MCPH* ($r = 0.543$, *adjusted* $p = 0.0$) and *MCPH** ($r = 0.565$, *adjusted* $p = 0.0$) show a strong and very highly significant correlation

with geography[233] (the distance matrices for *ASPM* and *MCPH* are in Annex 7.1). Again, the significant correlations are positive, suggesting that *genetic distances increase with spatial separation*; moreover, these correlations are higher than for linguistic features, suggesting a *stronger role of geography in shaping the genetic as opposed to the linguistic diversity*.

The (partial) Mantel correlations between the pair (*ASPM*, *MCPH*) and each linguistic feature are (Table 30):

Linguistic feature	*Correlation with (ASPM, MCPH)*		*Correlation with (ASPM, MCPH) controlling for geography*	
	r	*Adjusted p*	*r*	*Adjusted p*
Codas	0.478	0.0000	0.437	0.0000
NumNoun	0.382	0.0000	0.343	0.0025
Tone	0.333	0.0000	0.291	0.0025
WALSSylStr	0.243	0.0000	0.257	0.0025
GlotC	0.224	0.3322	0.205	0.3822
OnsetClust	0.116	0.4011	0.137	0.2332
VelarNasal	0.062	0.8100	-0.020	1.0000
NomLoc	0.086	1.0000	0.031	1.0000
AdjNoun	0.064	1.0000	0.054	1.0000
InterrPhr	0.071	1.0000	0.054	1.0000
RareC	0.106	1.0000	0.011	1.0000
Passive	0.075	1.0000	0.039	1.0000
VowelsCat	0.014	1.0000	0.015	1.0000
MorphImpv	0.042	1.0000	0.021	1.0000
ZeroCopula	0.027	1.0000	0.016	1.0000
ConsCat	0.017	1.0000	0.029	1.0000
CaseAffixes	0.014	1.0000	-0.004	1.0000
OVWO	0.002	1.0000	-0.099	1.0000
GenNoun	0.008	1.0000	0.009	1.0000
TenseAspect	-0.008	1.0000	-0.075	1.0000
Affixation	-0.021	1.0000	-0.080	1.0000

233Supporting again the missing data handling procedure for Africa.

Linguistic feature	*Correlation with (ASPM, MCPH)*		*Correlation with (ASPM, MCPH) controlling for geography*	
	r	*Adjusted p*	*r*	*Adjusted p*
FrontRdV	-0.028	1.0000	-0.036	1.0000
NumClassifiers	-0.063	1.0000	-0.207	1.0000
SVWO	-0.077	1.0000	-0.117	1.0000
AdposNP	-0.032	1.0000	-0.101	1.0000
UvularC	-0.041	1.0000	-0.014	1.0000

Table 30: Mantel correlations between the pair (*ASPM*, *MCPH*) and each linguistic feature individually without and with controlling for geography.
Gray: significant at the 0.05 level (Holm mcc).

Only 4 linguistic features have highly significant Mantel correlations with the pair (*ASPM*, *MCPH*): *Codas*, *NumNoun*, *Tone* and *WALSSylStr*, both before and after controlling for geography, and there seems to be a *substantial correlation not explained by geography* (paired t-test is non-significant: $t = 1.9736$, $df = 3$, $p = 0.1430$). These residual correlations range between 0.257 (*WALSSylStr*) and 0.437 (*Codas*), mean = 0.332. Thus, it can be concluded that *for these 4 linguistic features there is a significant correlation with (ASPM, MCPH) even when spatial distance has been factored out.*

The relationship between the pair (*ASPM*, *MCPH*) and the triplet[234] (*Codas*, *Tone*, *WALSSylStr*) is non-null: the distance matrices for this pair and triplet are reproduced in Figures 56 and 57, the Mantel correlation between (*ASPM*, *MCPH*) and geography is important and highly significant ($r = 0.5237$, $p < 0.01$), while the correlation between (*Codas*, *Tone*, *WALSSylStr*) and geography is low but also highly significant ($r = 0.1824$, $p < 0.01$). The correlation between (*ASPM*, *MCPH*) and (*Codas*, *Tone*, *WALSSylStr*) is important and highly significant ($r = 0.3884$, $p < 0.01$) and decreases only very slightly when controlling for geography ($r = 0.3497$, $p < 0.01$), suggesting that there exists a *correlation between these two genetic variants and the composite tone-syllable structure, which is not explained by spatial distances.*

234See Section 4.6.

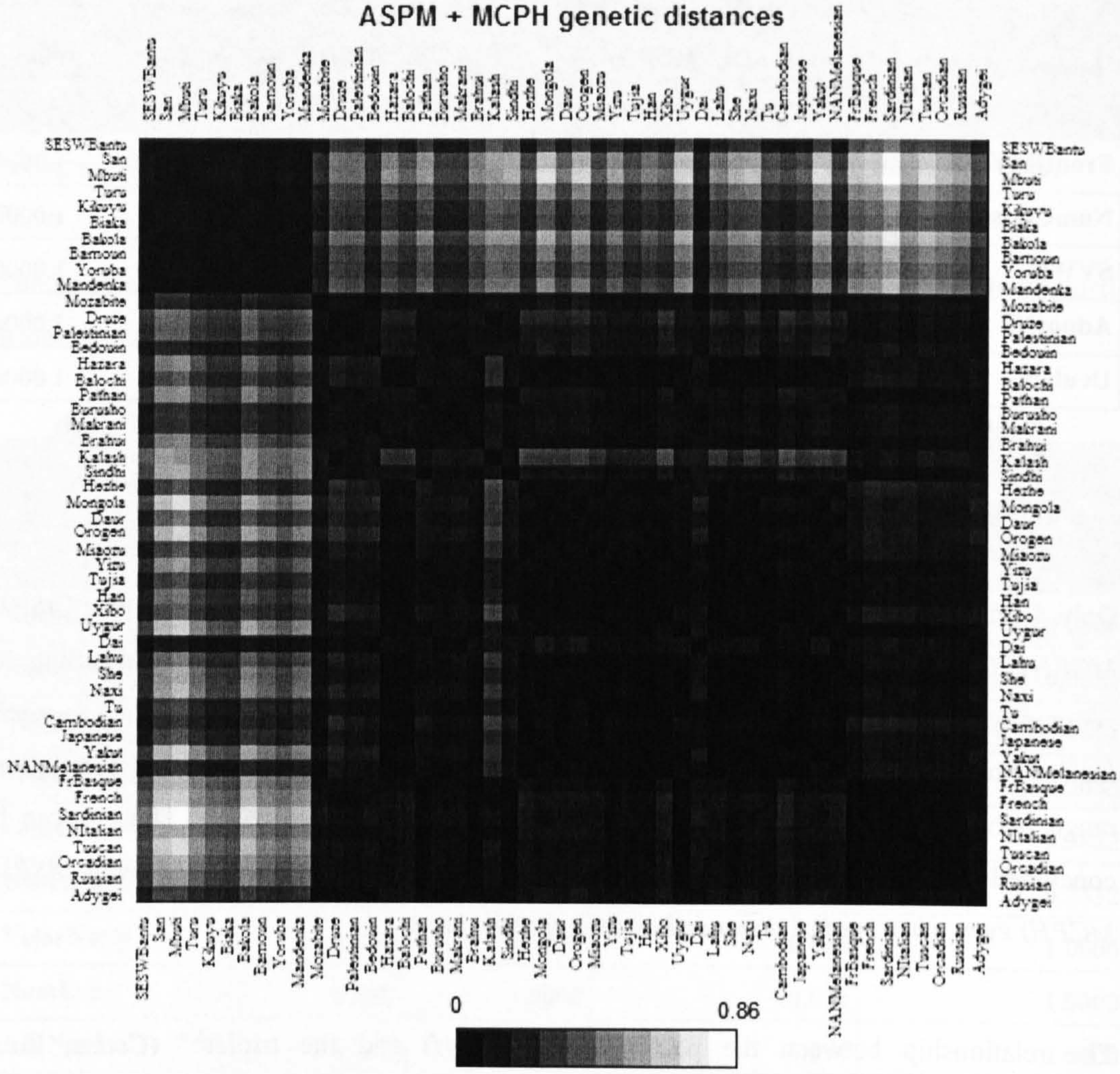

Figure 56: The (*ASPM*, *MCPH*) genetic distances (Nei's *D*) matrix: black (0) to white (0.86).

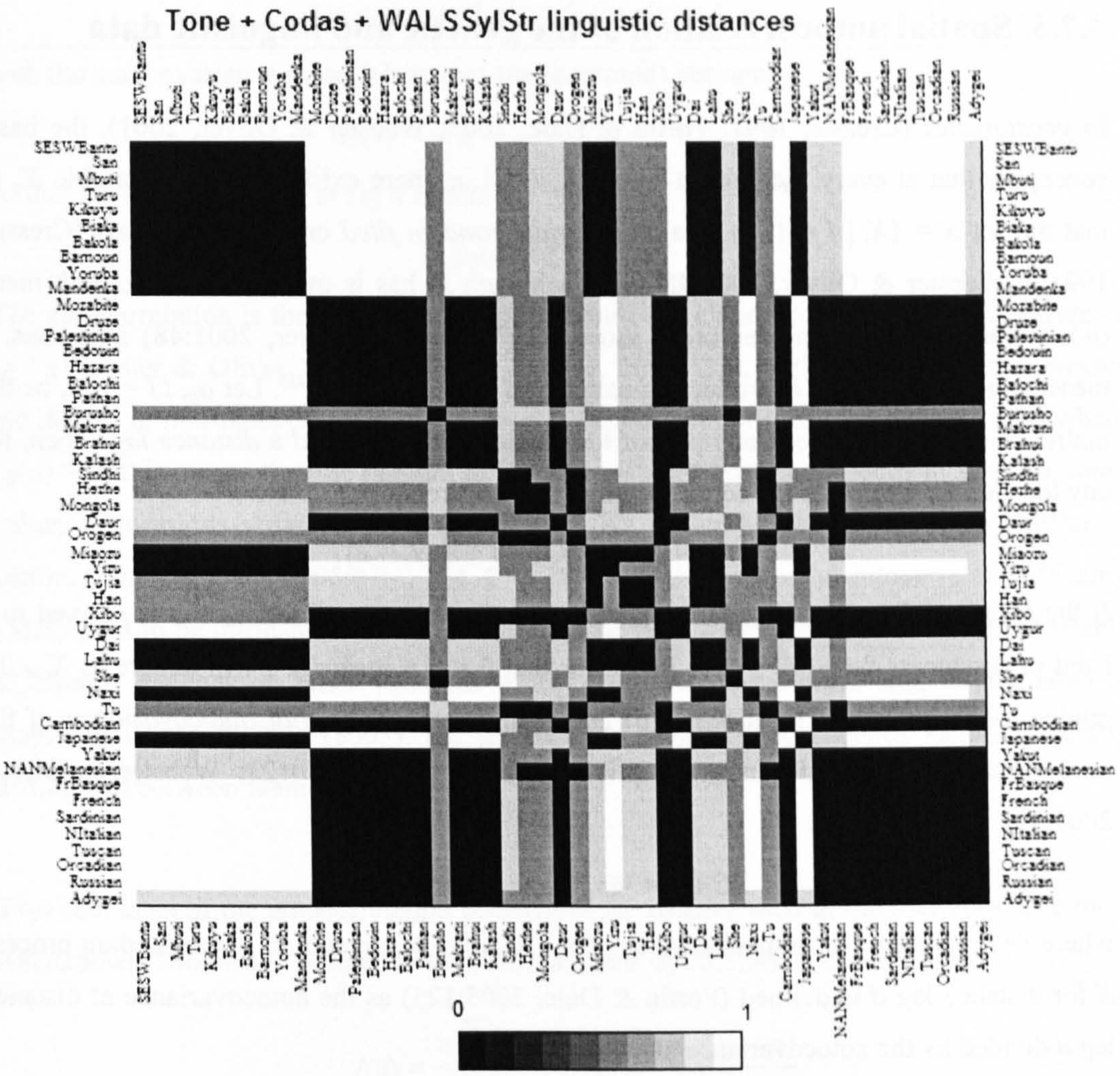

Figure 57: The (*Codas*, *Tone*, *WALSSylStr*) linguistic distances matrix: black (0) to white (1).

But no matter how useful, or popular, the (partial) Mantel test is, it does not capture the entire complexity of spatial relationships. For example,

> [w]hen we control for the effects of a third matrix, **C**, such as the Euclidean distances matrix among the sampling locations [or the land distances, in our case], *we are not controlling for the degree of spatial autocorrelation of the variables but only for the relative distances among the sampling locations*. [...] Furthermore, when the variables are strongly spatially autocorrelated, the restricted randomization (by rows and columns of the matrices) are no longer equally likely, so that *the significance of the partial Mantel test is not adequately evaluated* [...] (Fortin & Dale, 2005:151-152, *italics* mine),

so that different techniques must be used to study the other aspects of spatial dependency, like the spatial autocorrelation of the variables themselves.

4.7.3. Spatial autocorrelation of the genetic and linguistic data

In *geostatistics* (Cressie, 1991; Fortin & Dale, 2005; Webster & Oliver, 2001), the basic concept is that at every location in space, L_i, $i = 1..n$, there exists a random variable X_i, so that the set $X = \{X_i \mid i = 1..n\}$ is a *multivariate random filed* or *random process* (Cressie, 1991:8; Webster & Oliver, 2001:48-49). Each such X_i has is own distribution, with mean (μ_i), variance (σ^2_i) and higher order moments (Webster & Oliver, 2001:48) and, thus, a measurement at location L_i produces a realization of X_i, notated x_i[235]. Let d_{ij}, $i,j = 1..n$, be the matrix of distances between any pair of locations, L_i and L_j. and d a *distance lag*. Then, for any location L_i, we define the set of locations at distance d from L_i:

$$\Lambda_i(d) = \{L_j \mid i \neq j,\ 1 \leq j \leq n,\ d_{ij} = d\}.$$

If the locations L_i are not regularly spaced, then the exact condition $d_{ij} = d$ is relaxed to a band of distances, $d\text{-}\delta \leq d_{ij} \leq d+\delta$. It is clear that $0 < d \leq \max(d_{ij})$. I will denote as X_{i+d} the random variables at the locations $\Lambda_i(d)$ and, with these, the *spatial autocovariance* of the random process X for distance lag d is (Fortin & Dale, 2005:123; Webster & Oliver, 2001:51):

$$C_X(d) = \langle (x_i - \mu_i)\cdot(x_{i+d} - \mu_{i+d}) \rangle$$

where <•> represents the expected value. The *spatial autocorrelation* of the random process X for distance lag d is defined (Fortin & Dale, 2005:123) as the autocovariance at distance lag d divided by the autocovariance at lag 0:

$$\rho_X(d) = C_X(d) \,/\, C_X(0)$$

Unfortunately, in general, the solution of $\rho_X(d)$ is unavailable as we have only one realization of the random process X at each location L_i (Webster & Oliver, 2001:51), which impedes the estimation of the properties of the local random variables X_i. A way out is the assumption of *stationarity*, through which the distribution of the random process X is taken to have some spatially invariant properties (Webster & Oliver, 2001:52), irrespective of the absolute location and direction in space, such as the mean and variance (Fortin & Dale, 2005:11)[236]. Thus, assuming stationarity, the means and variances of the local random variables are the same:

235For a thorough theoretical treatment of the random spatial variables, see for example, Webster & Oliver (2001) or Cressie (1991, esp. pp. 7-26).

236For a detailed treatment and classification of types of stationarity, see Cressie (1991, esp. Section 2.3, pp. 52-58) and Webster & Oliver (2001:53-54).

$$\mu_i = \mu \text{ and } \sigma^2_i = \sigma^2, \text{ for all } i = 1..n$$

and, the autocovariance at lag 0 becomes the (common) variance:

$$C_X(0) = \langle (x_i - \mu_i)^2 \rangle = \langle (x_i - \mu)^2 \rangle = \sigma^2$$

so that the autocorrelation at lag d becomes:

$$\rho_X(d) = C_X(d) / \sigma^2$$

The autocorrelation is the equivalent of the standard correlation coefficient, ranging from -1 to 1 (Webster & Oliver, 2001:55; Fortin & Dale, 2005:123), with $\rho_X(0) = 1$, and expresses *the degree of non-independence of values at different locations*. Let's consider a case where $\rho_X(d) > 0$: this means that the random process X, assumed stationary, tends to have the same values at locations separated by a distance d. This tendency may be due to two different factors (Fortin & Dale, 2005:6-11, 124, 212-221): an *inherent* spatial autocorrelation (nearby values tend to be intrinsically more similar) and an *induced* spatial dependence (the variable of interest depends on other variables which show autocorrelation), and, as discussed at length in Fortin & Dale (2005), the spatial autocorrelation coefficients cannot generally distinguish between them.

Two estimates of the autocorrelation coefficient are usually used in the literature. The most well-known, Moran's I (Moran, 1948; Fortin & Dale, 2005:124), is defined as:

$$I(d) = \frac{n \cdot \sum_{i \neq j} w_{ij}(d) \cdot (x_i - \mu) \cdot (x_j - \mu)}{W(d) \cdot \sqrt{\sum_i (x_i - \mu)^2}}$$

where $w_{ij}(d)$ is the *distance class connectivity* (or *weight*) matrix (Fortin & Dale, 2005:124) and can be either the binary connection matrix ($w_{ij}(d) = 1$ if $L_j \in \Lambda_i(d)$, otherwise 0) or the numeric inverse distance matrix ($w_{ij}(d) = 1/d_{ij}$), while

$$W(d) = \sum_{i \neq j} w_{ij}(d)$$

It varies between -1 and 1, values close to 1 indicate strong positive autocorrelation, values close to -1 indicate strong negative autocorrelation, while values close to 0 indicate lack of autocorrelation[237]. It must be noted that $I(d)$ is the average value of spatial autocorrelation at distance d, in all directions, for the entire area, representing a *global isotopic average* (Fortin & Dale, 2005:125). $I(d)$ is potentially affected by outliers (Fortin & Dale,

237The expected value in the case of a total lack of autocorrelation is $-1/(n-1)$, which is very close to 0 for large n (Fortin & Dale, 2005:124): $\lim_{n \to \infty} -1/(n-1) = 0$. In our case, with n = 49, the expected value is –0.0208.

2005:125), and Geary's *c* (Geary, 1954; Fortin & Dale, 2005:126) was designed to avoid this pitfall (Fortin & Dale, 2005:126):

$$c(d) = \frac{(n\text{-}1)\cdot\sum_{i\neq j} w_{ij}(d)\cdot(x_i - x_j)^2}{2\cdot W(d)\cdot\sqrt{\sum_i (x_i - \mu)^2}}$$

Geary's *c* varies between 0 (strong positive autocorrelation) to 2 (strong negative autocorrelation), with values close to 1 indicating lack of autocorrelation. It is also potentially biased by outliers, but in a different manner than Moran's *I* (Fortin & Dale, 2005:126), so that it is better to use *both* coefficients simultaneously. For both Moran's *I* and Geary's *c*, if the inverse distance weight matrix ($w_{ij}(d) = 1/d_{ij}$) is used, there is no need for the distance lag *d*, and the computed autocorrelation coefficients *I* and *c* are global, reflecting the autocorrelation for the entire set of locations[238].

The actual values of the global *I* and *c* for all linguistic features and *ASPM* and *MCPH* are reproduced in Table 31, while summaries for the entire set of data are presented in Table 32.

Variable	*Moran's I*		*Geary's c*	
	statistic	*Adjusted p-value*	*statistic*	*Adjusted p-value*
MCPH	0.164	0.000	0.438	0.000
ASPM	0.178	0.000	0.634	0.000
NumNoun	0.147	0.000	0.708	0.000
AdjNoun	0.165	0.000	0.742	0.000
OVWO	0.128	0.000	0.736	0.000
RareC	0.146	0.000	0.618	0.000
Tone	0.121	0.000	0.718	0.000
WALSSylStr	0.152	0.000	0.719	0.000
VelarNasal	0.105	0.000	0.795	0.000
TenseAspect	0.136	0.000	0.691	0.038
Affixation	0.163	0.000	0.724	0.038
NumClassifiers	0.229	0.000	0.663	1.000
MorphImpv	0.132	0.000	0.808	1.000
Codas	0.172	0.000	0.829	1.000

238The actual implementation used was written in `R` (R Development Core Team, 2006) and is based on `R`'s `spdep` package, methods `moran` and `geary`. It also computes the *p*-value of the estimated autocorrelation coefficients using a randomization test (Edgington, 1987) which generates 1000 permutations of the values at each location.

Variable	*Moran's I*		*Geary's c*	
	statistic	*Adjusted p-value*	*statistic*	*Adjusted p-value*
AdposNP	0.089	0.028	0.815	0.038
CaseAffixes	0.125	0.028	0.772	0.038
GenNoun	0.089	0.216	0.914	1.000
ZeroCopula	0.076	0.528	0.935	1.000
NomLoc	0.077	1.000	0.913	1.000
UvularC	0.050	1.000	0.944	1.000
OnsetClust	0.080	1.000	1.069	1.000
VowelsCat	0.042	1.000	0.895	0.630
InterrPhr	0.053	1.000	0.968	1.000
ConsCat	0.029	1.000	0.939	1.000
Passive	-0.029	1.000	1.026	1.000
FrontRdV	0.022	1.000	0.959	1.000
GlotC	0.010	1.000	0.813	1.000
SVWO	-0.001	1.000	0.567	1.000

Table 31: The global autocorrelation estimators Moran's *I* and Geary's *c*.

The p-values were estimated using a randomization technique (Holm mcc). Dark gray: both autocorrelation coefficients are significant at the 0.05 level; light gray: only Moran's I is significant at the 0.05 level.

Set of variables	*Moran's I*			*Geary's c*			*Correlation*	*% significant at the 0.05 level*		
	Min	*Mean*	*Max*	*Min*	*Mean*	*Max*		*Moran's I & Geary's c*	*Moran's I only*	*Geary's c only*
Linguistic features	-0.029	0.096	0.229	0.567	0.818	1.069	-0.547	42.31%	11.54%	0.00%
genetic variants	-0.062	0.055	0.243	0.433	0.843	1.299	-0.767	7.93%	8.55%	2.03%

Table 32: Summary of the autocorrelation coefficients.

The correlations are computed as Pearson's r between the estimates of the autocorrelation coefficients (Moran's I and Geary's c). The percent of significant cases at the 0.05 level (Holm mcc) contains cases with both coefficients significant, those with only Moran's I significant and those with only Geary's c significant. Both correlations are significant at the 0.01 level.

In general, the two autocorrelation coefficients correlate strongly, reinforcing each other. 7.93% (78) of all genetic variants and an astonishing 42.3% (11) linguistic features show significant autocorrelations, suggesting that, *globally, the linguistic features have a greater tendency to be autocorrelated than the genetic variants*. 8.55% (84) of all genetic variants and 11.54% (3) linguistic features show autocorrelations significant only for Moran's *I* but not for Geary's *c*, while 2.03% (20) of all genetic variants and 0.0% (0) linguistic features show autocorrelations significant only for Geary's *c* but not for Moran's *I*, showing that these two estimates reflect differently the spatial structure of the data. The range of variation for these two coefficients is quite similar across domains (linguistic vs. genetic), with most of the cases showing no global spatial autocorrelation. Both *ASPM* and *MCPH* show a slight positive spatial autocorrelation, while *WALSSylStr* and *Tone* also show some spatial structure.

Another tool used in geostatistics for assessing the spatial structure of a variable is represented by the (*semi-*) *variogram*, which is the amount of variance in the data at different spatial lags (Fortin & Dale, 2005:132-138; Webster & Oliver, 2001:47-134; Cressie, 1991:58-104). It allows a condensed graphical representation of the spatial behaviour of the variable of interest, to which a theoretical variogram, from a selected set of models, can be fitted, representing the first step in hypothesizing an explanatory mechanism. For a given distance lag *d*, the *isotropic variogram* is defined by the *semi-variance function*[239]:

$$\gamma(d) = \frac{1}{2 \cdot n} \sum_i (x_i - x_{i+d})^2$$

Let δ be a *distance lag increment*, $0 \leq \delta \leq \max(d_{ij})$; then, for every distance lag $d_k = k \cdot \delta$, k □ □ so that $0 \leq d_k \leq \max(d_{ij})$, $\gamma(d_k)$ is computed and its graph against d_k represents the (semi-) variogram. In general, for non-regularly spaced locations, $\gamma(d_k)$ is computed for all locations distanced not by an exact distance lag d_k, but by the band of distances $(d_k - \delta/2, d_k + \delta/2]$. Thus, the variogram represents graphically *the mean variance existing between two locations separated by the current distance lag*, when the distance lag covers the entire space from 0 to the maximum distance between localities in steps of an atomic lag increment. It is important to highlight that the *y* axis is dimensionless and varies between 0 (no variance) to the maximum possible variance[240] and that the actual shape of the variogram depends on the

239For details on the computation of sample or experimental variograms, see, for example, Fortin & Dale (2005:132-134).

240In the case of binary variables (called *indicator variables* in geostatistics - Fortin & Dale, 2005:137), the maximum variance is 0.5.

scale (the distance lag δ), which must be chosen depending on the particularities of the study. Given that our sample is too small (see Fortin & Dale, 2005, esp. Chapter 5), a proper variographical study, including the fitting of theoretical variograms, must be postponed until better sampling will become available.

Figure 58 depicts the number of pairs of populations per distance lag, while Figures 59 - 61 present the variograms of all linguistic features and two genetic variants of interest. These variograms have a high diversity of shapes, suggesting very different spatial patterns. The two genetic variants, *ASPM* and *MCPH*, show relatively similar patterns[241], where dissimilarity increases with distance up to a maximum, but for large spatial scales, there is an increase in similarity. Some linguistic features (e.g., *MorphImpv*, *GenNoun*, *FrontRdV*, *ConsCat*, *Affixation*, *AdjNoun*) show an abrupt increase in variance, followed by a plateau, suggesting a very small spatial scale of interaction. Others (*ZeroCopula*, *WALSSylStr*, *VowelsCat*, *Tone*, *OVWO*, *OnsetClust*, *Codas*, *CaseAffixes*, *AdposNP*) show a gradual increase in variance until a (local) maximum is reached, followed by a decrease in variance at medium scales and again followed by an increase in variance for large scales, suggesting that different mechanisms work at different scales. For example, at small scales, contact and/or shared ancestry tends to produce low variance, while at medium scales languages assume essentially random values for these features, but the increased similarity at large scales is intriguing and suggests a series of hypothesis, including neighbor-inhibition-like processes[242] or ancient macro-areas of similarity[243] fractured by more recent processes[244]. Another pattern (*Passive*, *NumNoun*?) is represented by the monotonic increase in variance with the spatial lag, suggesting a mechanism based on divergence with distance, while some (*GlotC*, *InterrPhr*, *SVWO*, *TenseAspect*, *VelarNasal*) present a very rugged pattern, suggesting a low spatial dependence.

241 But on different scales: the maximum variance of ASPM is only 0.05072 while of MCPH is 4 times greater, 0.20864.

242 Whereby neighboring (at certain distance scales) regions tend to have different values from a limited set, conducing to sinusoidal patterns.

243 Due to, for example, ancient linguistic family spreads (akin to Nichol's spread zones) or to Dixonian equilibrium states.

244 Not necessarily of a different nature.

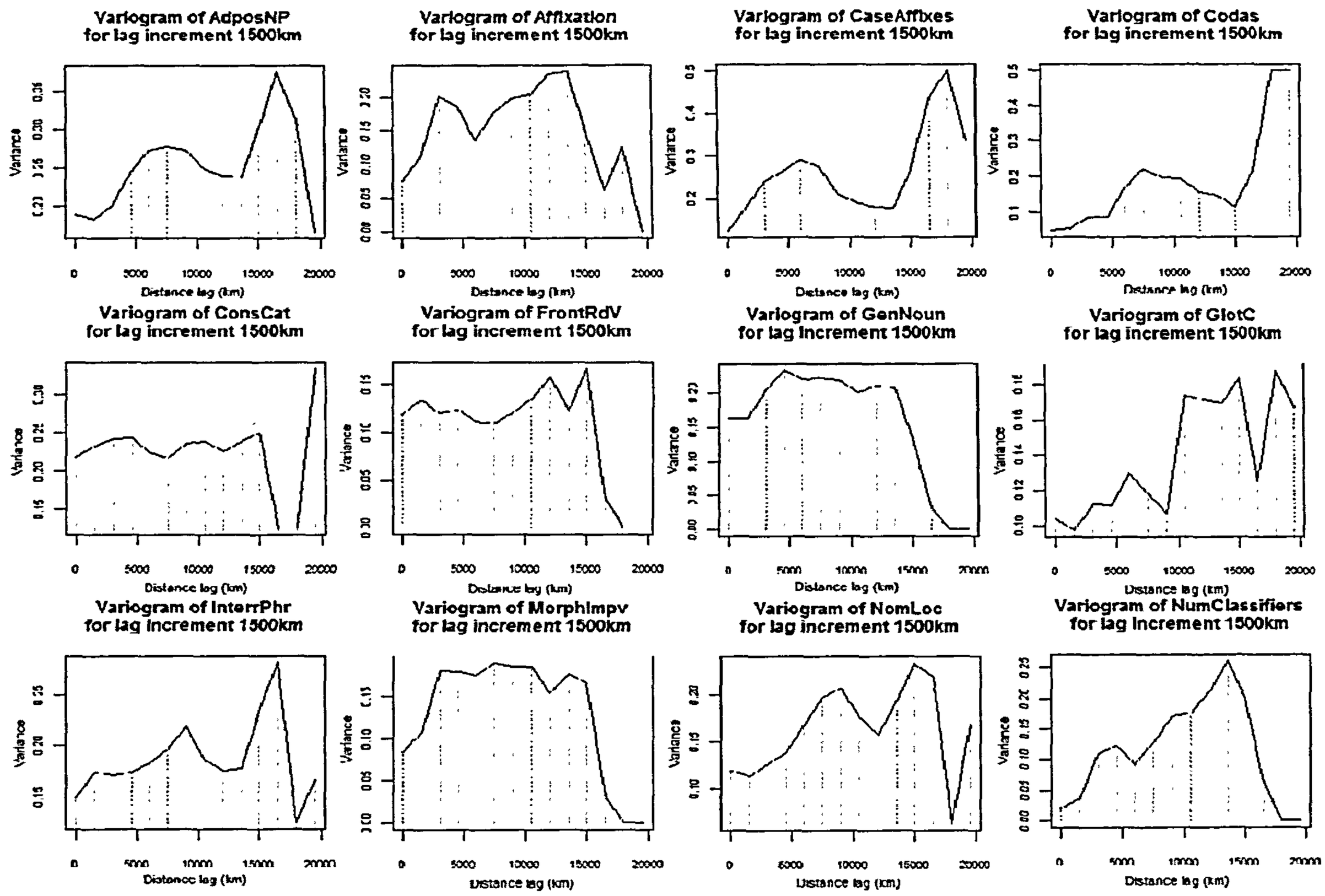

Figure 59: Variograms of *AdposNP*, *Affixation*, *CaseAffixes*, *Codas*, *ConsCat*, *FrontRdV*, *GenNoun*, *GlotC*, *InterrPhr*, *MorphImpv*, *NomLoc* and *NumClassifiers*.

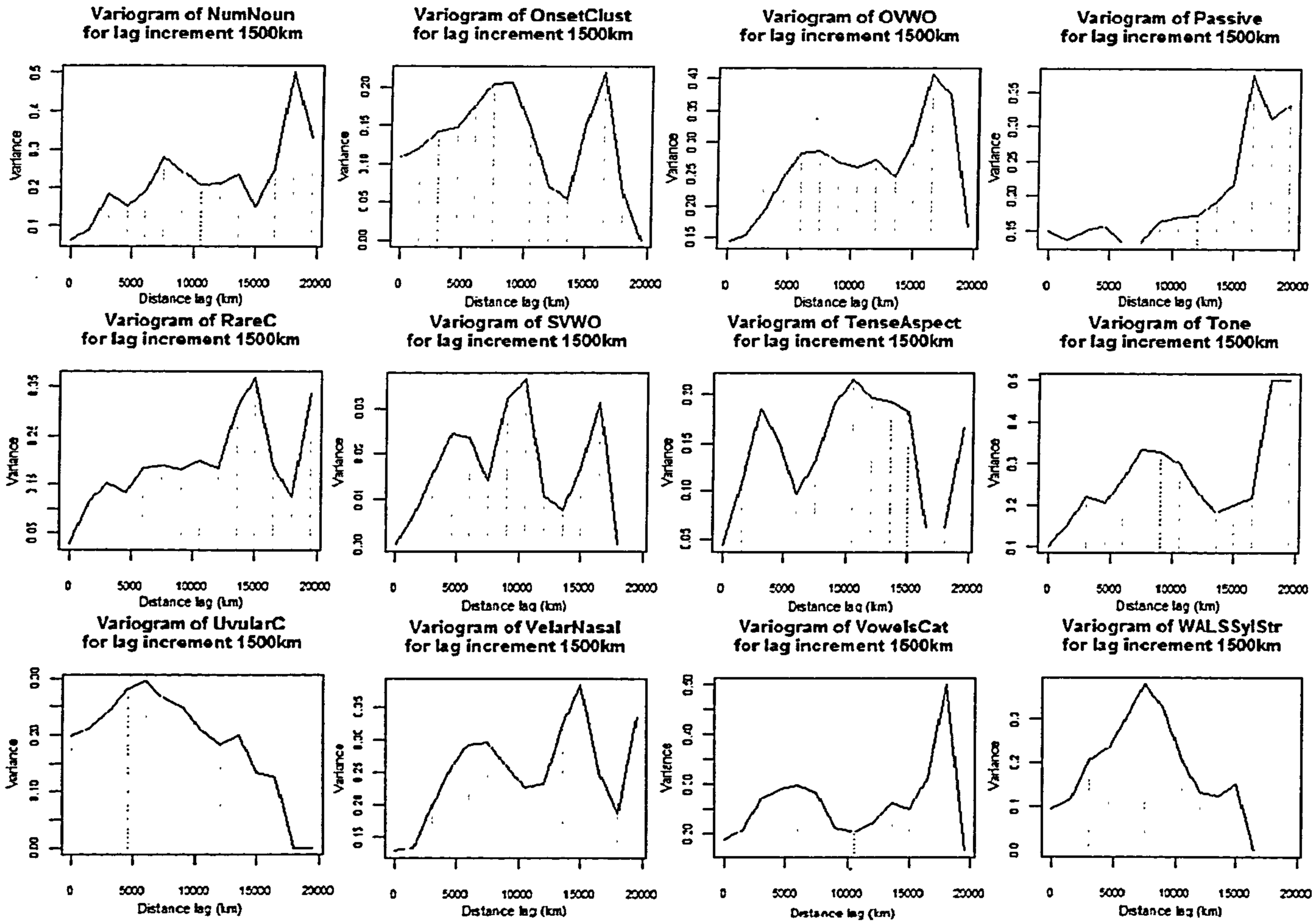

Figure 60: Variograms of *NumNoun*, *OnsetClust*, *OVWO*, *Passive*, *RareC*, *SVWO*, *TenseAspect*, *Tone*, *UvularC*, *VelarNasal*, *VowelsCat* and

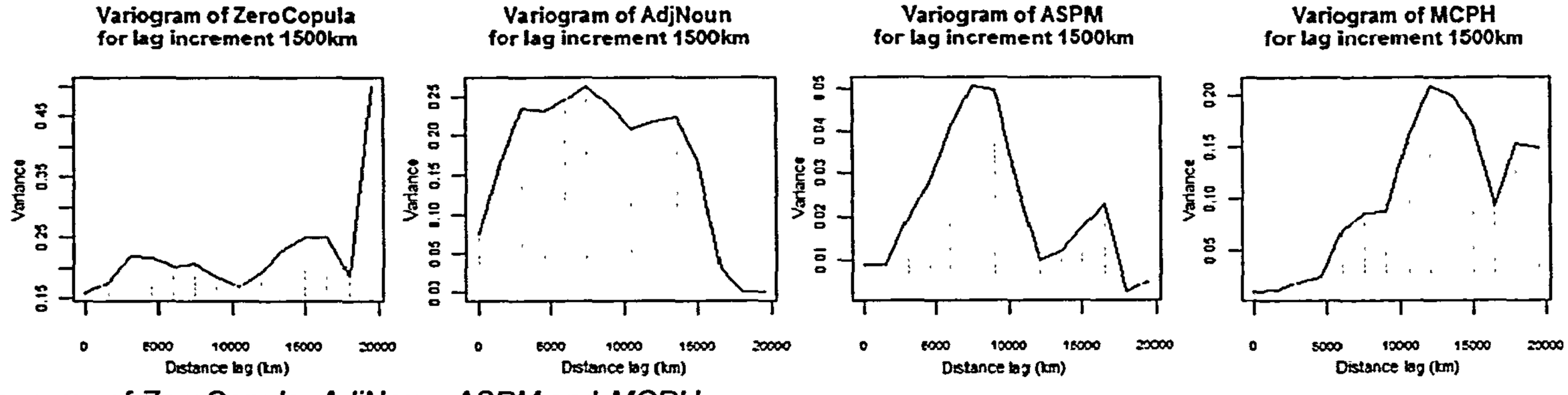

Figure 61: Variograms of *ZeroCopula*, *AdjNoun*, *ASPM* and *MCPH*.

Tone, *WALSSylStr* and *Codas* show the sinusoidal pattern:

Linguistic feature	***1st maximum***		***2nd minimum[245]***		***2nd maximum***	
	Distance lag	***Variance***	***Distance lag***	***Variance***	***Distance lag***	***Variance***
Tone	7500	0.333	13500	0.185	18000	0.500
WALSSylStr	7500	0.379	15000	0.150	-	-
Codas	7500	0.223	15000	0.117	18000	0.500

Table 33: Characteristics of the sinusoidal patterns of *Tone*, *WALSSylStr* and *Codas* for lag increment 1500km.

The values at spatial lags greater than 15,000km have low reliability, due to the very small number of population pairs. There is no 2nd maximum for WALSSylStr.

The 1st maximum occurs for all three linguistic features at 7500km: at this spatial scale, there are 321 pairs of populations[246] (Annex 7.2). The distance lag 7500 ± 1500km separates populations from different continents. The 2nd minimum (the 1st is the *nugget* at spatial lag 0), occurs at 13,500km (*Tone*) or 15,000 km (*WALSSylStr* and *Codas*). At 13,500km there are 65 pairs of populations (Annex 7.3), while at 15,000km, there are 30 pairs of populations (Annex 7.4). The 13,500 ± 1500km scale tends to oppose African with East and South Asian populations, neglecting Europe, while the 15,000 ± 1500km scale opposes Europeans to NAN Melanesians and Africans to East Asians. The 18,000km spatial lag separates only 8 populations from Sub-Saharan Africa with NAN Melanesian. *It can be concluded, thus, that the 7500km ± 1500km scale represents the maximum of linguistic dissimilarity on Tone and syllable structure, while the 13,500km – 15,000km scale, connecting populations from the extremes of the Old World, highlights a high similarity on these linguistic features.*

A very important caveat, potentially also affecting the interpretation of Moran's *I* and Geary's *c*, is that given the global nature of our data, the stationarity assumption might not hold and a more local approach might be needed. Nevertheless, it can be concluded that *some linguistic features and genetic variants show interesting autocorrelational spatial patterning*, but, for the moment, it is impossible to distinguish between various competing explanatory mechanisms.

245The 1st minimum occurs at distance lag 0 (the *nugget*).
246This is not the maximum number of pairs (339, reached at 6000km).

4.7.4. Genetic and linguistic boundaries

Another potentially very informative approach is to identify linguistic, genetic and geographic boundaries and to evaluate their correspondence, as this might shed light on the processes leading to linguistic and genetic differentiation. Given a set of populations and a distance matrix between them, the first step is to compute the *Delaunay triangulation* corresponding to the populations' geographical locations. For a set of points, P_1, P_2, ... P_n, this is the set of triangles constructed by joining triplets of points, P_i, P_j, P_k, so that the circumcircle of the triangle $\Delta P_i P_j P_k$ does not contain any other point from the set (Fortin & Dale, 2005:60-61; Okabe, Boots, Sugihara, 1992:72-76, 89-115). The Delaunay triangulation is closely related to the *Voronoi tessellation*, a concept very much used in spatial statistics (Fortin & Dale, 2005; Okabe, Boots, Sugihara, 1992)[247]. Intuitively, this triangulation captures the notion of *nearest neighbors in a set of geographical locations*.

The Delaunay triangulation for the 49 populations (no connection points) is represented in Figure 62. It must be pointed out that the following *boundary analysis* depends critically on the available sample, due to the detection of the neighboring populations through the Delaunay triangulation. Therefore, the analysis presented here is intended as a pilot study, illustrating this type approach, and exploring its potential usefulness to linguistic and genetic diversity problems.

247Okabe, Boots & Sugihara (1992) offer a very comprehensive but technical treatment.

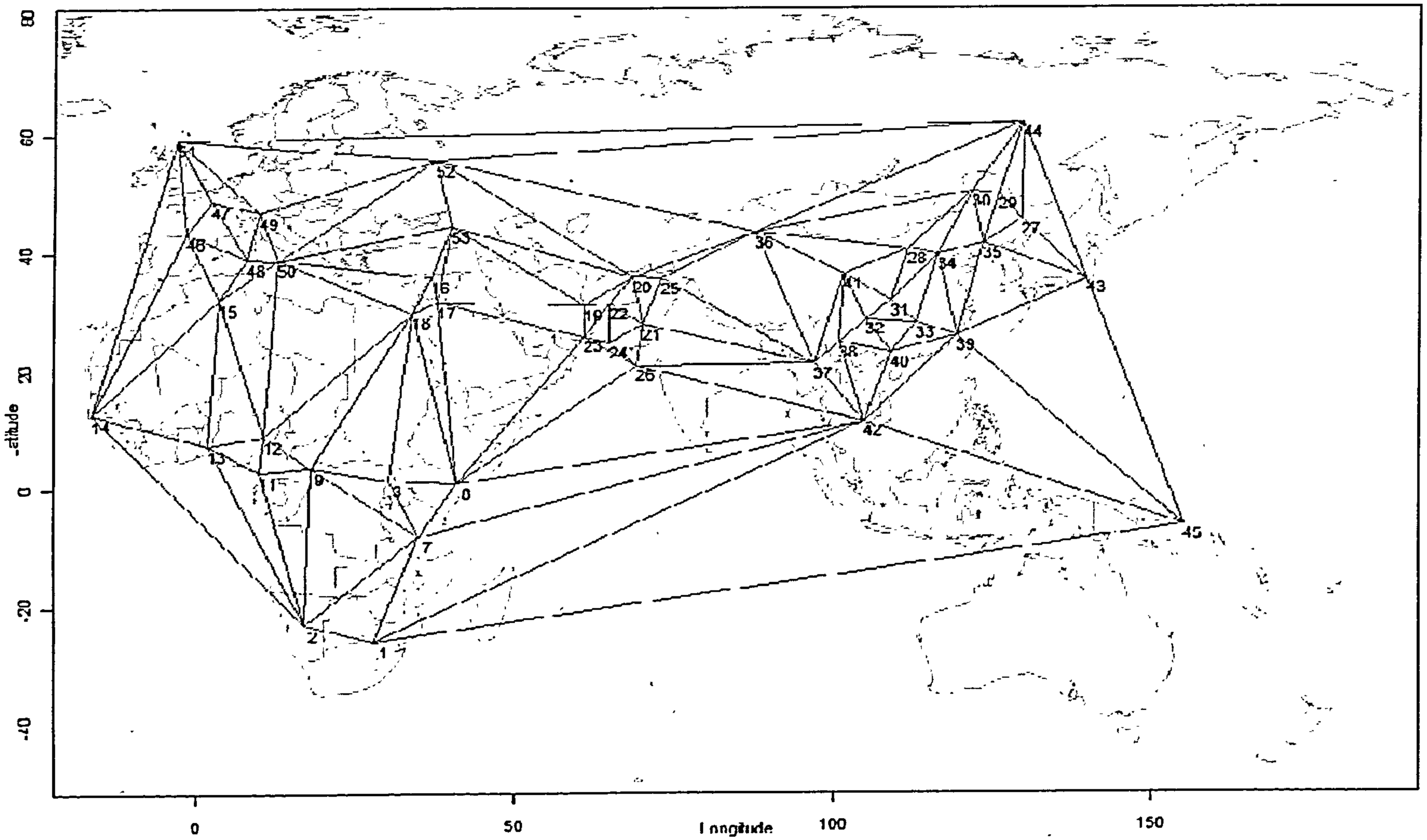

Figure 62: The Delaunay triangulation of the considered populations.

The population IDs are given in Figure 39. The geographical positions are approximate and have been adjusted for maximum clarity of the figure.

For a distance matrix, the Delaunay triangulation can be represented so that the width of the connections reflects the distance between neighboring populations, and the color of the connection can have two values: non-boundary connections and boundaries. In this context, *a boundary is defined as that distance between two neighboring populations greater than a given threshold value*[248]. Two methods of defining a threshold value were tested:

i. the threshold value is computed as *the top percent of the maximum distance* between two neighboring populations: if the maximum distance between any two neighbors is d_{max} and the threshold is τ ($0 < \tau < 1$), then the threshold value is $(1-\tau)d_{max}$;

ii. the threshold value is computed as *the value of the given topmost distances* between two neighboring populations: if there are n distances, then the $(\tau n)^{th}$ ($0 < \tau < 1$) topmost distance is used as the threshold value.

The Delaunay triangulations for the linguistic, genetic and land distances for both methods (i) and (ii) are represented in Annex 7.5, while Figures 63 – 65 reproduce only the maps for method (ii), threshold $\tau = .25$. The high-threshold cases ($\tau = .10$) produce too few boundaries, especially for genetic and land distances with method (i), so that only the cases with $\tau = .25$ and method (ii) will be analyzed.

248Boundary detection is a complex field in spatial statistics, as discussed, for example, by Fortin & Dale (2005:184-211). The approach used here is extremely simple.

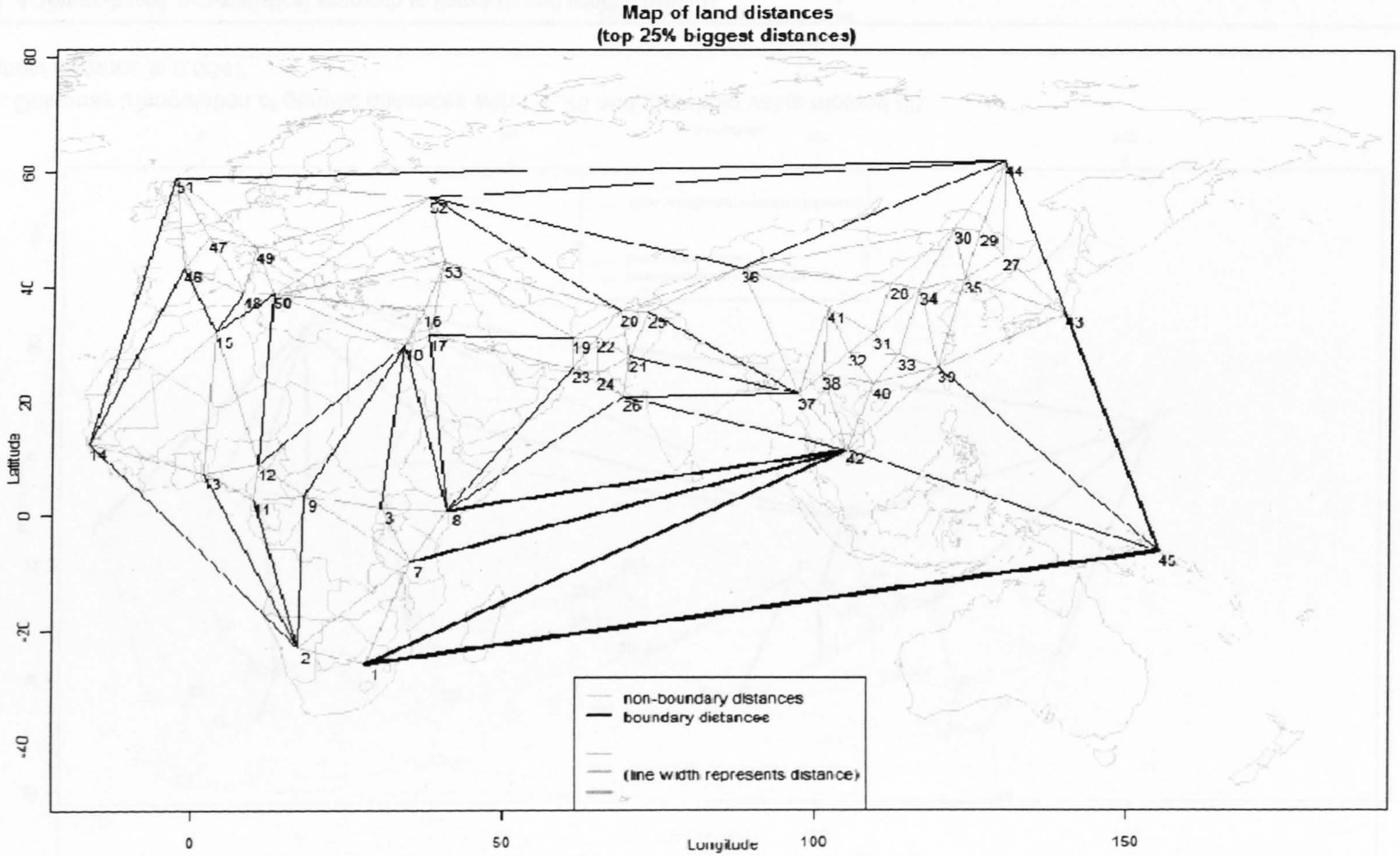

Figure 63: Delaunay triangulation of land distances with τ = .25 and threshold value method (ii).

The threshold distance is 2974.62

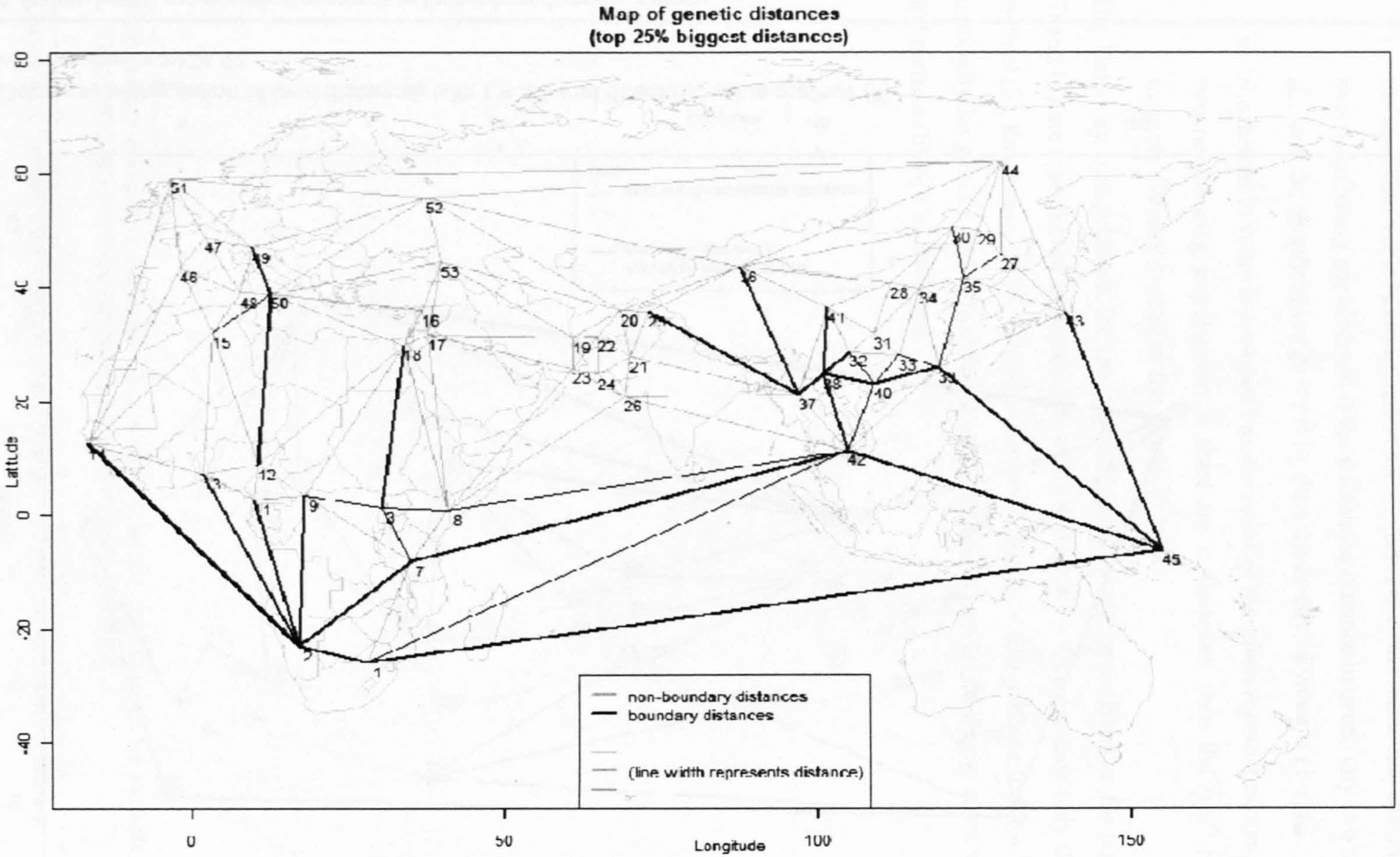

Figure 64: Delaunay triangulation of genetic distances with τ = .25 and threshold value method (ii).
The threshold distance is 0.0547.

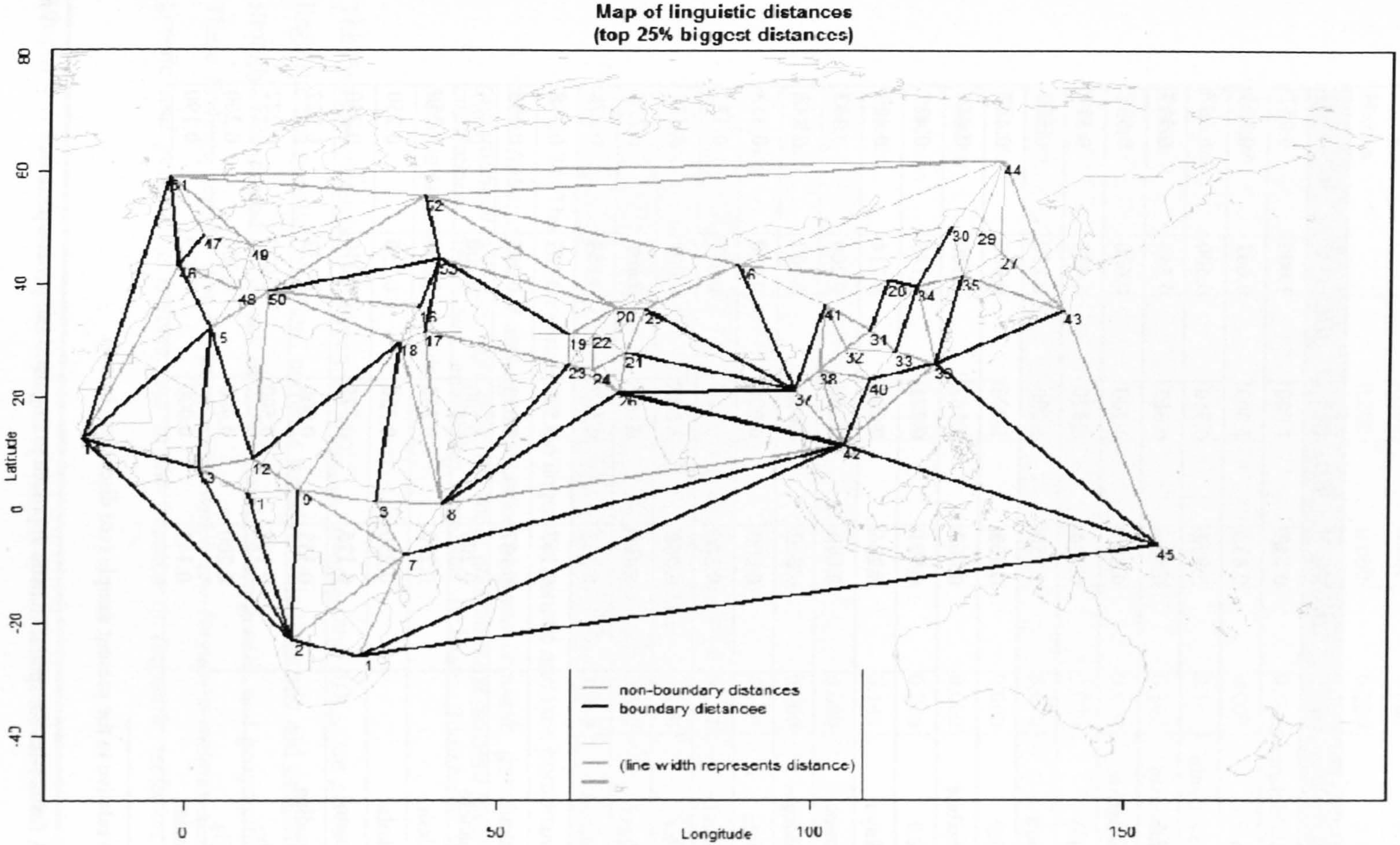

Figure 65: Delaunay triangulation of linguistic distances with τ = .25 and threshold value method (ii).
The threshold distance is 0.6202.

A useful statistic is given by the ratio of the number of boundaries originating from a given population to the total number of Delaunay lines originating from that population, the *boundary density*, $0 \leq BD(\text{population}) \leq 1$: a population with high *BD* is more distinctive than one with a low *BD*, and a *BD* of 1 means that there are boundaries all around the node. Thus, *BD* can be conceptualized as *a measure of "isolation"*[249] (Table 34).

Population	$BD_{Linguistic}$	$BD_{Genetic}$	BD_{Land}	BD_{Avg}
NANMelanesian	0.750	1.000	1.000	0.917
San	0.833	1.000	0.667	0.833
SESWBantu	0.750	0.750	0.500	0.667
Cambodian	0.556	0.667	0.556	0.593
Mandenka	0.800	0.200	0.600	0.533
Dai	0.714	0.429	0.429	0.524
Mbuti	0.250	1.000	0.250	0.500
Naxi	0.400	1.000	0.000	0.467
Mozabite	0.667	0.167	0.500	0.444
She	0.571	0.571	0.143	0.429
Kikuyu	0.286	0.286	0.714	0.429
Yoruba	0.600	0.200	0.200	0.333
FrBasque	*0.600*	*0.000*	*0.400*	*0.333*
Sindhi	0.500	0.000	0.500	0.333
Bedouin	0.286	0.143	0.571	0.333
Lahu	***0.000***	***1.000***	***0.000***	***0.333***
Adygei	***0.833***	***0.000***	***0.000***	***0.278***
Bamoun	0.333	0.167	0.333	0.278
Turu	0.167	0.500	0.167	0.278
Biaka	0.167	0.333	0.333	0.278
Japanese	0.200	0.200	0.400	0.267
Kalash	0.250	0.250	0.250	0.250
Bakola	0.250	0.250	0.250	0.250
Tuscan	0.125	0.375	0.250	0.250
Orcadian	0.333	0.000	0.333	0.222
Xibo	0.167	0.500	0.000	0.222
Tujia	0.200	0.400	0.000	0.200
Russian	0.143	0.000	0.429	0.190

249Isolation relative to the present sample (see discussion below).

Population	$BD_{Linguistic}$	$BD_{Genetic}$	BD_{Land}	BD_{Avg}
Yakut	***0.000***	***0.000***	***0.571***	***0.190***
Han	0.500	0.000	0.000	0.167
Palestinian	0.167	0.000	0.333	0.167
Uygur	0.125	0.125	0.250	0.167
Mongola	0.400	0.000	0.000	0.133
Hazara	0.200	0.000	0.200	0.133
Yizu	0.000	0.400	0.000	0.133
Tu	0.167	0.167	0.000	0.111
Orogen	0.167	0.167	0.000	0.111
Makrani	0.167	0.000	0.167	0.111
Pathan	0.167	0.000	0.167	0.111
French	0.250	0.000	0.000	0.083
Druze	0.250	0.000	0.000	0.083
Hezhen	0.000	0.250	0.000	0.083
Miaozu	0.200	0.000	0.000	0.067
NItalian	0.000	0.200	0.000	0.067
Sardinian	0.000	0.000	0.200	0.067
Balochi	0.000	0.000	0.143	0.048
Daur	0.000	0.000	0.000	0.000
Brahui	0.000	0.000	0.000	0.000
Burusho	*0.000*	*0.000*	*0.000*	*0.000*
Mean	**0.296**	**0.259**	**0.241**	**0.265**
Median	**0.200**	**0.167**	**0.200**	**0.222**
Min	**0.000**	**0.000**	**0.000**	**0.000**
Max	**0.833**	**1.000**	**1.000**	**0.917**

Table 34: The *BD* measures for the linguistic, genetic and land borders.

***Bold italic**: cases of striking dissociation between linguistic, genetic and geographical isolation. Light gray: the most and the least isolated (overall) populations. Italic: two interesting linguistic isolates: Burushaski and Basque (see text for details).*

The Pearson correlations between the $BD_{Linguistic}$, $BD_{Genetic}$ and BD_{Land} are given in Table 35: linguistic and genetic "isolation" correlate weakly, while linguistic and geographic correlate strongly. Unexpected is the lack of correlation between genetic and geographic "isolation". These findings suggest that *the populations which are linguistic isolates also tend to be genetic and geographic isolates, but in the absence of linguistic isolation, geographic*

isolation seems not to develop into genetic isolation[250]. An interesting case is represented by Adygei, which is a strong linguistic isolate ($BD_{Linguistic}$(Adygei) = 0.833) while genetically and geographically it is not isolated ($BD_{Genetic}$(Adygei) = BD_{Land}(Adygei) = 0). A different pattern is shown by Lahu, which is a very strong genetic isolate ($BD_{Genetic}$(Lahu) = 1.000) but not geographically and linguistically ($BD_{Linguistic}$(Lahu) = BD_{Land}(Lahu) = 0), while Yakut is geographically isolated (BD_{Land}(Yakut) = 0.571), but not linguistically or genetically ($BD_{Linguistic}$(Yakut) = $BD_{Genetic}$(Yakut) = 0). The cases of the linguistic isolates Basque, and especially Burusho, show the decoupling of historical linguistic isolation from areal linguistic, genetic and geographic isolation. Thus, *linguistic, genetic and geographical isolation do not coincide perfectly* and there are striking cases of dissociation. The most isolated (linguistically, genetically and geographically[251]) populations are NANMelanesian, San, SESWBantu, Cambodian, Mandenka, Dai and Mbuti[252], while the least isolated[253] are Daur, Brahui and Burusho.

Correlations	$BD_{Genetic}$	BD_{Land}
$BD_{Linguistic}$	0.307*	0.509**
$BD_{Genetic}$		0.261

Table 35: Pearson's correlations between *BD* for linguistic, genetic and land boundaries.

*Significance levels: *: significant at $p < 0.05$, **: significant at $p < 0.01$, otherwise, statistically non-significant.*

In order to compare the boundaries across types of distances in a "global" manner (as opposed to the "local" approach offered by *BD*), a *boundary matrix*, *BM*, was generated for each distance used. The rows and columns of the *BM* are populations and an entry of this matrix is 1 if the corresponding pair of populations are connected through a boundary, and 0 otherwise. Thus, for example, $BM_{Linguistic}$(SESWBantu, San) = 1 while $BM_{Linguistic}$(SESWBantu, Turu) = 0. A measure of the correspondences between two boundary matrices, BM_1 and BM_2,

---

250This very interesting result needs better sampling in order to insure its generality, as it might be due simply to the current sample's small size.

251$BD_{Avg} > 0.5$.

252These finds depend crucially on the available sample, as better sampling modifies the neighborhood of the populations, drastically altering the boundary landscape. A good example is provided by Mbuti, which in our sample appears as linguistically isolate ($BD_{Linguistic}$(Mbuti) = 0.714), while, in fact, it is known that they represent a case of language shift under the pressure of the neighboring Bantu speaking agriculturalists. This case highlights again the crucial importance of a good sample.

253$BD_{Avg} = 0.0$.

is represented by two times[254] the ratio of boundaries shared by the two matrices to the total number of boundaries in the two matrices, and denoted $SB(BM_1, BM_2)$ (*shared boundaries*). As $SB(BM_1, BM_2)$ approaches 1, the more boundaries are shared by the two boundary matrices, while an $SB(BM_1, BM_2)$ close to 0 represents two boundary matrices containing different boundaries. The values of *SB* for this sample are[255]:

Shared boundaries (SB)	$\boldsymbol{BM_{Genetic}}$	$\boldsymbol{BM_{Land}}$
$\boldsymbol{BM_{Linguistic}}$	0.421	0.500
$\boldsymbol{BM_{Genetic}}$		0.429

Table 36: The ratio of shared boundaries to total number of boundaries (SB) between two boundary matrices computed using different distance measures.

Approximately half the boundaries are shared between the three modalities: linguistic, genetic and geographic, suggesting again that there is *a strong connection between linguistic and genetic differentiation and geographical remoteness*. A measure of the relationship between linguistic and genetic boundaries when controlling for geographical boundaries is given by the *partial SB*, defined for three matrices BM_1, BM_2 and BM_3, as the number of shared boundaries of BM_1 and BM_2 minus the number of shared boundaries between all three matrices divided by the total number of boundaries of BM_1 and BM_2. For our sample,

$$SB_{Partial}(\boldsymbol{BM_{Linguistic}}, \boldsymbol{BM_{Genetic}}; \boldsymbol{BM_{Land}}) = 0.1316,$$

showing that there is *a small residual set of shared linguistic and genetic boundaries when geography has been controlled*[256]. Thus, it can be concluded that *even if geographical boundaries explain most of the shared linguistic and genetic boundaries, there also exist some other processes responsible for a limited number of such shared boundaries*.

The boundary analysis presented in this section is intended as a pilot study into this very complex but potentially extremely relevant area. Its conclusions are tentative in the extreme, given their sensitivity to the sample used and their requirement for as close spatially as possible samples in order to detect real abrupt changes. Nevertheless, it proves promising,

---

254In order to normalize this measure to the [0,1] interval.

255These values are very close the the Mantel correlations between the same matrices, which are: $r(\boldsymbol{BM_{Linguistic}},\boldsymbol{BM_{Genetic}}) = 0.4032$, $r(\boldsymbol{BM_{Linguistic}},\boldsymbol{BM_{Land}}) = 0.4850$ and $r(\boldsymbol{BM_{Genetic}},\boldsymbol{BM_{Land}}) = 0.4110$, all significant at the $p < 0.01$ level. This coincidence is explained by the correlation's formula in the binary case.

256The partial Mantel correlation is $r(\boldsymbol{BM_{Linguistic}}, \boldsymbol{BM_{Genetic}}; \boldsymbol{BM_{Land}}) = 0.2557$, significant at the $p < 0.01$ level.

and further study is warranted. Some first steps towards a better analysis of linguistic, genetic and geographical borders will be the usage of larger sample sets and better techniques for border detection, based on the triangulation-wombling technique (Fortin & Dale, 2005:196-197). Also, very important will be the addition to the simple land distance between localities of ecological (e.g., deserts, rain forest, temperate steppe) and topographic boundaries (e.g., major mountain chains, rivers or narrow land bridges), given that these are potentially powerful shapers of linguistic and genetic diversity.

4.8. Controlling for history: historical linguistics, genes and linguistic features in a spatial context

Another potentially very powerful explanatory factor of linguistic (and genetic) diversity is reflected by the history of the languages. Historical linguistics offers a principled approach to explaining patterns of linguistic diversity through common descent and differentiation (McMahon &McMahon, 2005; Mallory, 1991). Therefore, it is important to integrate this dimension into our approach. The distribution of the linguistic families of the 26 languages is given in Figure 66 below. Most of languages belong to Indo-European (24.5%), Altaic[257] (16.3%), Niger-Congo (16.3%), Sino-Tibetan (10.2%) and Afro-Asiatic (8.2%).

In order to asses the impact of sharing the same linguistic family on the differentiation of the populations, all the possible pairs of populations ($49^2 = 2401$) were classified in two groups: the *shared linguistic family group* (*SLFG*) and the *different linguistic family group* (*DLFG*), based on the linguistic family of the languages spoken: a pair of populations, P_1 and P_2, speaking languages L_1 and L_2, belong into the *SLFG* if and only if L_1 and L_2 are from the same linguistic family, otherwise P_1 and P_2 belong into the *DLFG*. Thus, *SLFG* and *DLFG* partition the set of all unique population pairs (49*(49-1)/2 = 1176) in two disjoint classes, containing 139 and 1037 pairs, respectively. Two-samples t-tests were performed for the linguistic, genetic and land distances between the *SLFG* and *DLFG* in order to asses the impact of shared linguistic family on the linguistic and genetic similarity between populations (Table 37), which shows that *pairs of populations speaking languages from the same linguistic family, irrespective of the specific linguistic family, tend to cluster in space*

257See the discussion about the status of Altaic in Section 3.2.4.2. In this context, its acception is that given by Gordon (2005), which does not include, for example, Japanese or Korean.

(land distances), and to be overall more similar genetically and linguistically than pairs of populations speaking languages from different families. Also, for most linguistic features taken individually, languages from the same family tend to be more similar.

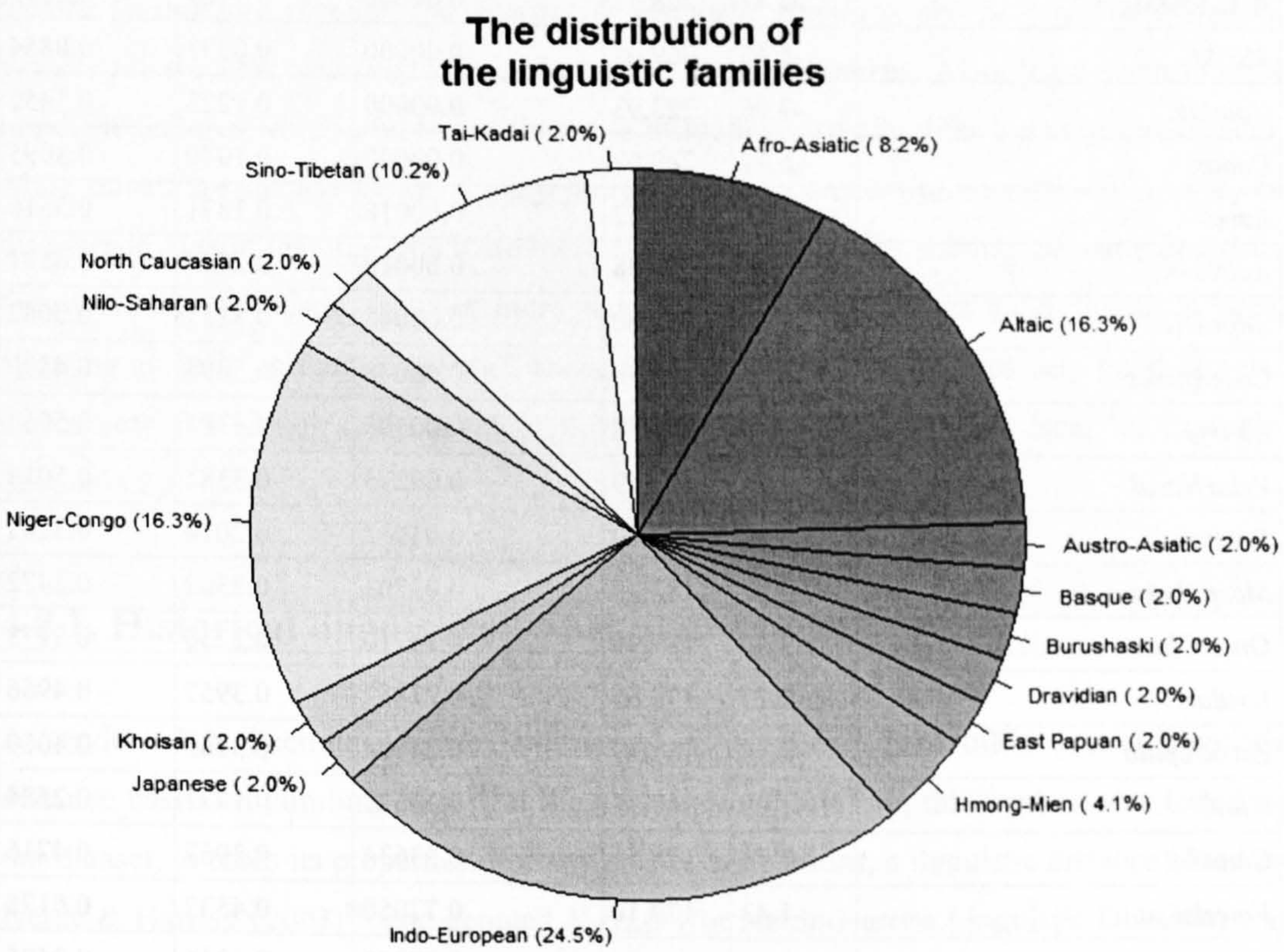

Figure 66: The (alphabetical) distribution of the language families of the considered languages.

Distance	*t-test*	*df*	*Adjusted p-value*	$mean_{SLFG}$	$mean_{DLFG}$
Tone	-36.26	1036.00	0.00000	0.0000	0.5593
ASPM & MCPH	-24.14	1105.64	0.00000	0.0253	0.1742
Land	-21.26	285.57	0.00000	2686.2310	6898.9860
MCPH	-21.03	1058.69	0.00000	0.0149	0.3372
NumClassifiers	-19.96	1036.00	0.00000	0.0000	0.2777
NumNoun	-19.15	548.96	0.00000	0.0288	0.4301
Codas, Tone & WALSSylStr	-15.87	221.03	0.00000	0.1352	0.5530
Affixation	-13.98	359.40	0.00000	0.0504	0.3857
All genetic variants	-13.80	228.11	0.00000	0.0313	0.0549
All linguistic features (*IPWS*)	-12.50	159.76	0.00000	0.4018	0.5616

Distance	*t-test*	*df*	*Adjusted p-value*	$mean_{SLFG}$	$mean_{DLFG}$
All linguistic features (*EWS*)[258]	-12.34	157.64	0.00000	0.4409	0.6146
All linguistic features *(DPWS*)	-12.02	157.03	0.00000	0.4543	0.6310
TenseAspect	-9.73	279.08	0.00000	0.0719	0.3288
WALSSylStr	-9.42	208.31	0.00000	0.1727	0.5092
ASPM	-8.45	269.04	0.00000	0.0333	0.0854
NomLoc	-7.06	223.93	0.00000	0.1223	0.3452
Codas	-6.71	229.02	0.00000	0.1079	0.3095
RareC	-4.51	197.52	0.00018	0.1871	0.3510
AdjNoun	-4.51	187.08	0.00018	0.2734	0.4581
AdposNP	-4.31	182.60	0.00048	0.3237	0.5082
CaseAffixes	-4.27	183.76	0.00048	0.3094	0.4899
OVWO	-3.90	181.58	0.00196	0.3381	0.5063
VelarNasal	-3.78	181.59	0.00273	0.3381	0.5014
Passive	-3.20	191.40	0.01932	0.2014	0.3202
MorphImpv	-3.02	188.29	0.03201	0.2302	0.3472
OnsetClust	-2.46	187.09	0.14710	0.2230	0.3173
UvularC	-2.27	178.65	0.21852	0.3957	0.4966
ZeroCopula	-2.18	181.59	0.24704	0.3165	0.4089
GlotC	-1.99	187.52	0.33635	0.1871	0.2584
ConsCat	-1.71	178.51	0.53634	0.3957	0.4716
VowelsCat	-1.43	177.10	0.77050	0.4532	0.5178
SVWO	-1.37	212.25	0.77050	0.0216	0.0405
GenNoun	-0.57	177.65	1.00000	0.3885	0.4137
FrontRdV	-0.41	178.69	1.00000	0.2302	0.2459
InterrPhr	-0.39	177.63	1.00000	0.3525	0.3693

Table 37: Two samples t-test for various distance measures between *SLFG* and *DLFG*.

Gray: significant at the $p < 0.05$ level (Holm mcc).

The last 10 features also have non-significant Mantel correlations with geography (Table 29), suggesting that these features are too labile, both historically and geographically, to carry any meaningful information. *Languages sharing a linguistic family tend to very strongly cluster spatially*, clustering which probably explains the highly significant differences concerning the genetic data inside and between families. Also, *Tone and the two*

258Given the similarity between the three weighting schemes, only EWS will be used in the following.

genetic variants combined (ASPM, MCPH) are extremely neatly distinguished between the two classes of populations. Nevertheless, a very important caveat against putting too much weight on the interpretation of these results is given by the scarcity and non-systematicity of the available samples, but even with this, it seems clear that *moderately deep historical factors (manifested through the sharing of linguistic families) are powerful explanatory devices for the observed linguistic and genetic diversity patterns*. Also, there seem to exist linguistic features which tend to remain stable inside linguistic families but very across them, while others are far too labile. A very promising direction for future research is to try to discover linguistic features which are more similar for languages sharing the same linguistic family versus linguistic features more similar between languages in a linguistic area, allowing an early assessment of such areas, but such a program requires very good samples and a very fine detail at the level of the chosen linguistic features[259] (see, for example, Section 3.2).

4.8.1. Historical linguistically-based distances

As reviewed in Section 3.2, the language-genes literature quite often uses a linguistic distance based on climbing historical linguistic trees. Therefore, this method was tested on our dataset, to asses its properties. For any pair of populations, a linguistic distance based on Nettle & Harriss' (2003)[260] and denoted *N-HLD* (for Nettle-Harriss Linguistic Distance) was computed, using the linguistic classification given in the *Ethnologue* (Gordon, 2005): *1*, same language; *2*, languages in the same branch of a family; *3*, languages in different branches of same family; or *4*, languages not demonstrably related. The resulting distances matrix is reproduced in Figure 67 and the Mantel correlations between *N-HLD* and the other distances used in this study (following the methodology in Poloni *et al.*, 1997 and Rosser *et al.*, 2000) are reproduced in Table 38.

259For example, *Tone* considered as a unitary linguistic features might prove too coarse for such an approach and it might be necessary to sub-analyze it into its various forms (e.g, lexical, morpho-syntactic, etc.).

260Compatible also with Poloni *et al.* (1997).

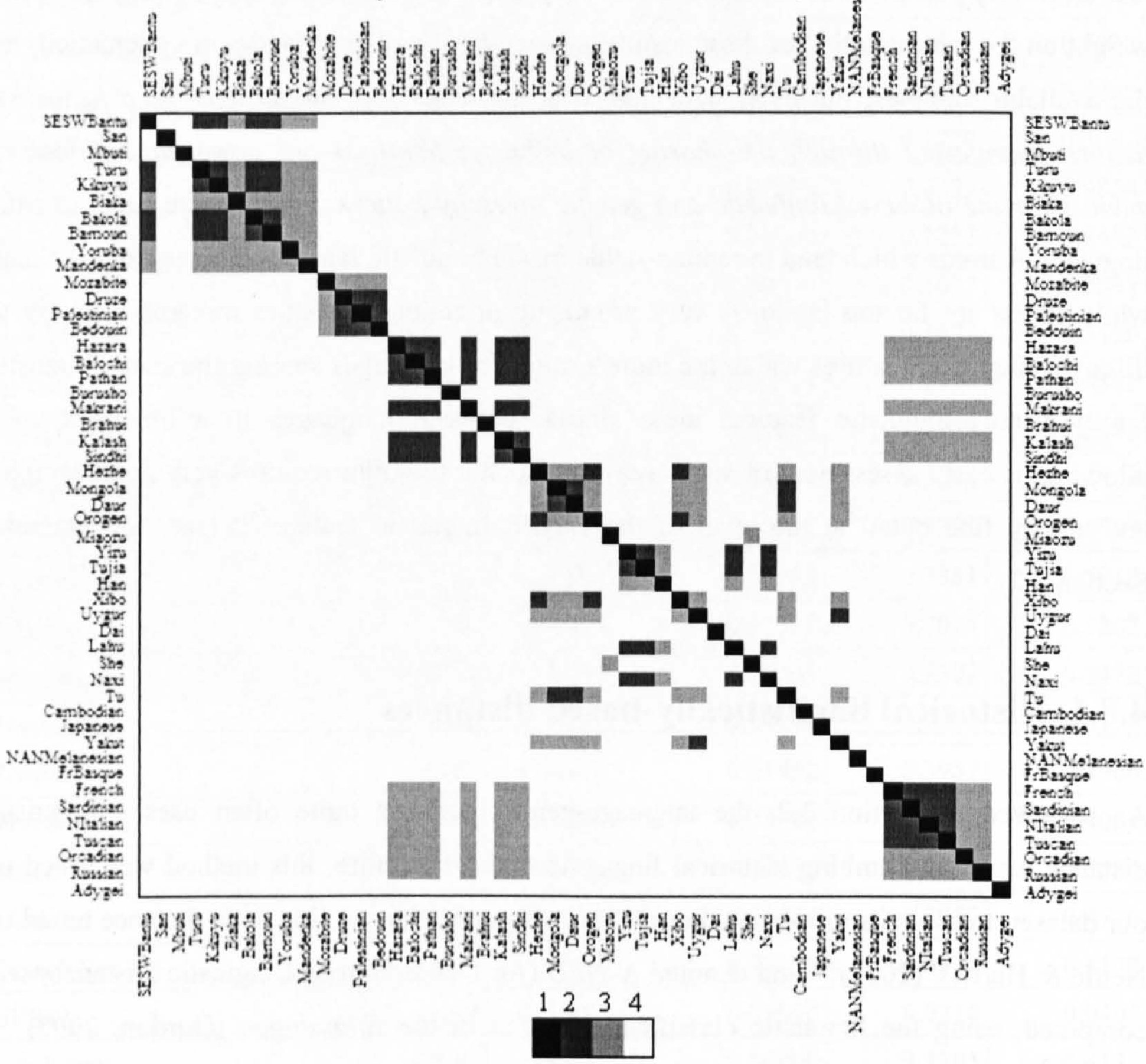

Figure 67: Linguistic distances between populations computed using Nettle & Harriss's (2003) method and the Ethnologue linguistic classification (Gordon, 2005).

1, same language; 2, languages in the same branch of a family; 3, languages in different branches of same family; or 4, languages not demonstrably related.

Distances	*Mantel's r*	*Adjusted p-value:*
Land	0.382	0.0006
Genetic (all markers)	0.277	0.0006
Linguistic (all features)	0.445	0.0006
Genetic controlled for land	0.104	0.0308
Linguistic controlled for land	0.380	0.0006
Linguistic vs. land controlling for *N-HLD*	0.137	0.0124

*T*able 38: The Mantel (partial) correlations between *N-HLD* and other types of distances used in this study.

Light gray: Mantel partial correlation between linguistic (features) and land distances when controlled for N-HLD. All significant at the 0.05 level (Holm mcc).

The correlation of *N-HLD* with geography is notable ($r = 0.3817$, $p < 0.01$), confirming Poloni *et al.*'s (1997:1018) findings, while its correlation with genetics is slightly lower ($r = 0.2772$, $p < 0.01$). The correlation with the linguistic features distance ($r = 0.4447$, $p < 0.01$) confirms that there is a strong historical component in linguistic features differentiation. The partial Mantel correlation between *N-HLD* and genetics when controlling for geography is low ($r = 0.1041$, $p = 0.0308$), suggesting (again) that *the correlation between linguistic (family) distribution and genetics is mostly explained by geography*, confirming previous conclusions (e.g., Poloni *et al.*, 1997; Rosser *et al.*, 2000; Section 3.2.4.6). The partial Mantel correlation between linguistic (features) distance and *N-HLD*, when controlling for geography is still important ($r = 0.3797$, $p < 0.01$), but slightly lower that the non-geographically controlled correlation ($r = 0.4447$), suggesting that *historical linguistic processes (linguistic family) are an important factor in determining overall linguistic features diversity, besides geographical proximity (language contact)*. In the same vein, the partial Mantel correlation between linguistic (features) distance and geography, when controlling for *N-HLD*, is low but significant ($r = 0.1372$, $p = 0.0124$), representing *the influence of language contact on linguistic (features) diversity besides shared linguistic ancestry*.

It is possible to control *simultaneously* for geography and historical linguistics through second-order partial Mantel correlations[261]. The zero-, first- and second-order Mantel

261 Where $r_{AB.XY} = (r_{AB.X} - r_{AY.X} \bullet r_{BY.X})/\sqrt{(1-r_{AY.X}^2)(1-r_{BY.X}^2)}$, with $r_{AB.X}$, $r_{AY.X}$ and $r_{BY.X}$ the first-order partial correlations; p is computed in the same way as for first-order partial Mantel correlations (through random permutations). The procedure was implemented in R on the model of `vegan`'s `mantel.partial`.

correlations for linguistic distances (all features), genetic distances (all markers), controlling for land and *N-HLD* distances are in Table 39 below:

Zero-order		***First-order (land)***		***First-order (N-HLD)***		***Second-order***	
r	***Adjusted p***	***r***	***Adjusted p***	***r***	***Adjusted p***	***r***	***Adjusted p***
0.162	**0.048**	0.021	0.865	0.045	0.865	-0.020	0.865

Table 39: Zero-, first- and second-order partial Mantel correlations between linguistic distances (all features) and genetic distances (all markers), when controlling for land and *N-HLD* distances.

Gray, bold: significant at the 0.05 level (Holm mcc).

and for genetic distances based only on *ASPM* and *MCPH*:

Linguistic feature(s)	***Zero-order***		***First-order (land)***		***First-order N-HLD***		***Second-order***	
	r	***Adjusted p***	***r***	***Adjusted p***	***r***	***Adjusted p***	***r***	***Adjusted p***
Codas	0.478	0.000	0.437	0.000	0.464	0.000	0.435	0.000
NumNoun	0.382	0.000	0.343	0.000	0.339	0.000	0.336	0.000
Tone	0.333	0.000	0.291	0.000	0.271	0.000	0.283	0.000
WALSSylStr	0.243	0.000	0.257	0.000	0.200	0.005	0.248	0.000
GlotC	0.224	0.320	0.205	0.337	0.222	0.361	0.205	0.306
OnsetClust	0.116	0.405	0.137	0.239	0.105	0.649	0.133	0.306
VelarNasal	0.062	0.853	-0.020	1.000	0.040	1.000	-0.023	1.000
NomLoc	0.086	0.986	0.031	1.000	0.052	1.000	0.023	1.000
RareC	0.106	1.000	0.011	1.000	0.077	1.000	0.007	1.000
Passive	0.075	1.000	0.039	1.000	0.056	1.000	0.035	1.000
InterrPhr	0.071	1.000	0.054	1.000	0.061	1.000	0.051	1.000
AdjNoun	0.064	1.000	0.054	1.000	0.029	1.000	0.044	1.000
MorphImpv	0.042	1.000	0.021	1.000	0.022	1.000	0.016	1.000
ZeroCopula	0.027	1.000	0.016	1.000	0.011	1.000	0.011	1.000
ConsCat	0.017	1.000	0.029	1.000	0.007	1.000	0.025	1.000
VowelsCat	0.014	1.000	0.015	1.000	0.001	1.000	0.011	1.000
CaseAffixes	0.014	1.000	-0.004	1.000	-0.020	1.000	-0.013	1.000
GenNoun	0.008	1.000	0.009	1.000	-0.002	1.000	0.006	1.000
OVWO	0.002	1.000	-0.099	1.000	-0.041	1.000	-0.107	1.000
TenseAspect	-0.008	1.000	-0.075	1.000	-0.058	1.000	-0.088	1.000

Linguistic feature(s)	*Zero-order*		*First-order (land)*		*First-order N-HLD*		*Second-order*	
	r	*Adjusted p*	*r*	*Adjusted p*	*r*	*Adjusted p*	*r*	*Adjusted p*
Affixation	-0.021	1.000	-0.080	1.000	-0.083	1.000	-0.097	1.000
FrontRdV	-0.028	1.000	-0.036	1.000	-0.034	1.000	-0.037	1.000
AdposNP	-0.032	1.000	-0.101	1.000	-0.074	1.000	-0.111	1.000
UvularC	-0.041	1.000	-0.014	1.000	-0.063	1.000	-0.022	1.000
NumClassifiers	-0.063	1.000	-0.207	1.000	-0.119	1.000	-0.218	1.000
SVWO	-0.077	1.000	-0.117	1.000	-0.090	1.000	-0.120	1.000
All	*0.027*	*1.000*	*0.016*	*1.000*	*0.011*	*1.000*	*0.011*	*1.000*

Table 40: Zero-, first- and second-order partial Mantel correlations between linguistic distances (each feature separately and all together) and genetic distances (*ASPM* & *MCPH* only), when controlling for geography (land distance) and history (*N-HL*D).

Gray: correlations significant at the 0.05 level (Holm's mcc). Last row (italic): all features together.

It can be seen that *Tone* still correlates with *ASPM* & *MCPH* even after simultaneously controlling for geography and history (*N-HLD* seems to be a suppressor variable for this correlation). Thus, this methodology seems able, in principle, to disentangle the contributions of historical linguistic and areal factors in shaping the linguistic diversity, but more work is needed.

The method applied by Nettle & Harriss (2003) to the study of genes-languages correlations is interesting but has some potential problems (Section 3.2.4.6). Its thorough application to this dataset is described in Annex 4 and the overall conclusion is that its usage in geographical studies of genetic and linguistic relationships is not warranted.

4.9. The relationship between ASPM, MCPH and Tone

As specified in Section 4.1, the *a priori* hypothesis is that there is a non-null relationship between *ASPM*, *MCPH* and *Tone*. It can be concluded, following the attempts to falsify it (Sections 4.6-4.8), that:

1. There is a statistically significant relationship between *ASPM* and *Tone* and *MCPH* and *Tone*, separately. This relationship also falls in the top 5% strongest in the entire

empirical sample of linguistic features and genetic variants.

2. There is also a statistically significant relationship between the pair (*ASPM, MCPH*) and *Tone*. This relationship also falls in the top 5% strongest in the entire empirical sample of linguistic features and genetic variants.
3. This relationship also holds when controlling for geography and history, simultaneously.
4. *ASPM, MCPH* and *Tone* show a strong spatial autocorrelational structure and all three tend to be much more similar inside linguistic families than across them (but this could be due to spatial clustering).

These points definitely reject the null hypothesis of no relationship between them, and this relationship is also very strong compared to all the relationships between the other linguistic features and genetic variants available. Therefore, it can be safely concluded that *there is correlation between the frequency of ASPM-D and MCPH-D haplogroups in a population and the probability that the language(s) spoken by that population will use tone contrasts, correlation not entirely explained by geographical proximity or common descent.*

From the scatter plot of *Tone* vs *ASPM* and *MCPH* (Figure 68) and the logistic regression coefficients (Table 25), results that:

- *low* frequencies of both *ASPM-D* and *MCPH-D* haplogroups are associated with the *presence* of tonal distinctions,
- *high* frequencies of both *ASPM-D* and *MCPH-D* haplogroups are associated with the *absence* of tonal distinctions, while
- *low* frequency of *ASPM-D* and *high* frequency of *MCPH-D* haplogroups are associated with an equal probability (10:11) of presence or absence of tone distinctions.

(There are no cases of high frequency of *ASPM-D* and low frequency of *MCPH-D*). Thus, the hypothesis can be further refined to state that *the lack of both derived haplogroups from a population is associated with the probable usage of tone distinctions, while the increase in their frequencies is associated with a linguistic trajectory trough coexistence of tonal and non-tonal systems towards the full dominance of non-tonal linguistic systems.*

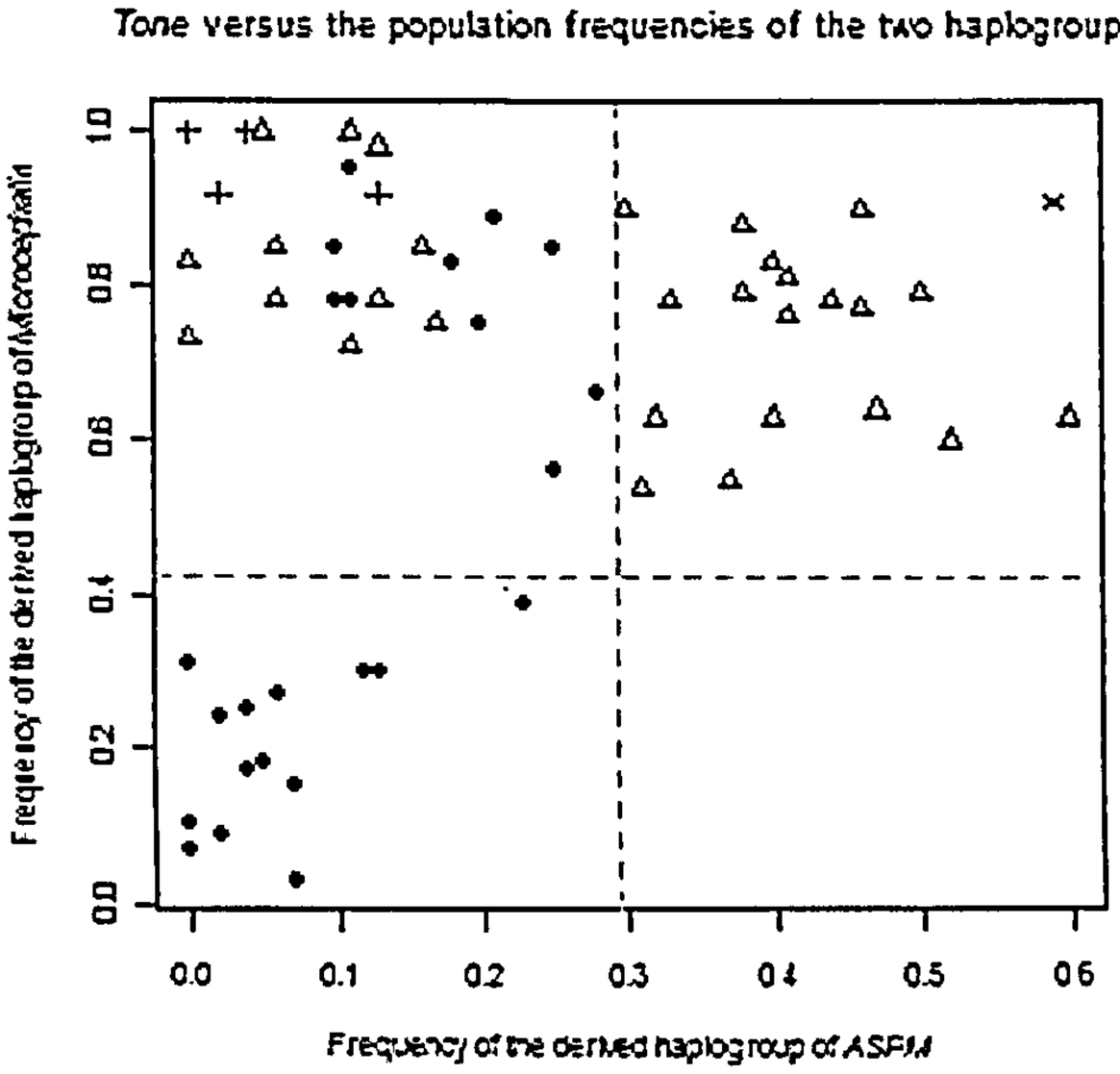

Figure 68: Scatter plot of *Tone* vs *ASPM* and *MCPH*.

The horizontal axis represents the frequency of ASPM-D, while the vertical axis represents the frequency of MCPH-D. Solid circles represent tonal languages, open triangles non-tonal languages, the crosses represent the American populations and the "X" represents the Papuan population. Gray dashed lines correspond to 0.239 ASPM-D and 0.425 MCPH-D. See text for details.

A tentative confirmation of this refined hypothesis is offered by the American populations: low frequencies of *ASPM-D* and high frequencies of *MCPH-D* are associated with a mixture of tonal and non-tonal languages. In the Papuan case the interpretation is more problematic. More specifically, the extremely high frequency of *ASPM-D* (59.4%, very close to the maximum of 60% - Kalash), and very divergent from its immediate neighbor, NAN Melanesian (11.1%), and more remote south-east Asian neighbors (Cambodian, 0%; Miaozu, 10% and She, 21.4%), casts an important doubt on the non-contamination of the Papuan sample with European genes. In fact, the Papuan sample is very much like the European samples, in what concerns the frequency of both *ASPM-D* and *MCPH-D*, but more extreme (t-tests, *ASPM*: $t = -8.3005$, $df = 5$, *adjusted* $p = 0.000829$, and *MCPH*: $t = -3.8718$, $df = 5$, *adjusted* $p = 0.011740$), possibly suggesting a founder effect from a small European admixing group. Therefore, until more controlled samples from the New Guinean highlands are available, nothing can be inferred about them. Another very important case is represented by Australia, with its seemingly absolutely non-tonal languages over the entire continent,

but, unfortunately, no genetic data are available[262]. Moreover, due to the recent history of the Australian aborigines, one must be very cautious about such samples concerning potentially selectively non-neutral genes.

Statistical analyses alone are notoriously incapable of inferring causal relationships from correlations and this is painfully true in our case. Therefore, we can only formulate explanatory hypotheses and try to falsify them with the current data and/or propose future studies targeted at this:

- ***Pure chance***. This is a plausible explanation for the observed pattern, but its probability is low.
- ***Same type of mechanisms***. It is possible that the mechanisms shaping the genetic and (feature) linguistic diversity are similar on a conceptual level, involving both vertical (common descent) and horizontal (contact) processes, and a set of common constraints (geography, ecology, history). Thus, it is possible that some combinations of parameters will produce coherent patterns, detected by these techniques. Unfortunately, it is not possible at this stage to accurately quantify this probability, as it requires specific computational and mathematical modeling studies, but it can still be said that it probably is not a fully satisfactory explanation, given the residual correlations remaining after controlling for geography and linguistic history.
- ***Deep demographic processes***. It seems plausible that some correlations between neutral genetic variants, reflecting demographic processes, and some linguistic features, are due to these linguistic features being *stable through time*. Possible mechanisms of stability include differential substratum effects in language replacements (Ostler, 2005) and conservatism in language change. Apparently, good candidates are those features correlating with many genetic variants and showing a non-random spatial patterning (from our data, such linguistic features could be *Codas*, *Tone*, *NumNoun* and *NumClassifiers*). It must be highlighted that these are just candidates and further study is required, but it seems plausible that different linguistic features have different stabilities through time and that detecting those in the right-hand tail of the distribution is possible. If so, something akin to Nichols'

262Concerning specifically Australia, the original team was considering obtaining samples in October 2005 (Bruce T. Lahn, *pc*), but no newer information was available to me at the time of this writing (September 2006).

(1992) program might prove feasible, whereby the most stable linguistic features and the genetic variants correlating with them, could shed light on very ancient demographic processes (expansions, replacements and bottlenecks). It could offer, thus, a complementary and exploratory side to the more rigorous but sensitive comparative method (Renfrew, McMahon & Trask, 2000; McMahon & McMahon, 2005).

- ***Causal links between ASPM, MCPH and Tone (non-spurious correlation)***. The previous explanation assumes that the correlation between genetic variants and linguistic features rests on a common demographic history of change and stability. But *ASPM* and *MCPH* seem not to be neutral, and, thus, their pattern does not reflect past demography, but specific selective pressures. Therefore, those linguistic features correlating specifically with them become good candidates for a non-mediated causal determinism (of course, pure chance could play a role, but its probability is rather low). There are four linguistic features correlating with both *ASPM* and *MCPH*: *Codas*, *WALSSylStr*, *NumNoun* and *Tone*, even after geography and history have been controlled for. These four linguistic features have a similar non-random spatial structure (autocorrelations and variograms) and all of them are more similar inside linguistic families than across them. Moreover, if we consider that *Codas* and *WALSSylStr* probably reflect the same subtending factor, *syllable structure*, and that *Tone* and *syllable structure* are, in fact, a reflection of a deeper layer of *sequential/parallel linguistic processing* (D.R. Ladd, *pc*), then we are left with only two correlations involving *ASPM* and *MCPH*. *It can be hypothesized, thus, that there is a direct (but complex) causal connection between the frequency of these derived haplogroups (ASPM-D and MCPH-D) in a population and the probability that tonal contrasts will be used (as a manifestation of the subtending sequential/parallel processing strategies)*[263].

The detailed discussion of the general theory of non-spurious correlations, its assumptions, implications and impact, will be treated in Chapter 5. Until then, it must be highlighted, briefly, that this non-spurious correlation, if confirmed, does not imply that tonal/parallel

263The case of NumNoun is very intriguing. It could be due to chance (word-order generally seems very labile) or there could be a real (but not yet understood) connection between tone-syllable structure and numeral-noun word order. The inspection of the appropriate maps in Haspelmath, Dryer, Gil & Comrie (2005) seems to suggest that a test case is represented by Australia, where Tone is patently absent, but NumNoun has both possible values.

languages are "more primitive" (whatever that means) or that tonal-language speakers are "less evolved" than speakers of more sequential languages. There is a series of plausible generic mechanisms bridging the gap between the frequency of *ASPM-D* and *MCHP-D* in a population and the probability of linguistically using tone contrasts:

1. high frequencies of these derived haplogroups might make possible the development of new, sequential, structures, which took over the pre-existing parallel structures, or
2. high frequencies of these derived haplogroups might make hard to acquire parallel structures, forcing the speakers to use more heavily sequential structures.

It is highly improbable that the positive selection on *ASPM-D* and *MCPH-D* is due to their linguistic effects, and it is much more plausible that these are neutral by-products of the genes' main functions connected to brain growth and development (see Chapter 5 for a thorough discussion of these issues). Thus, the linguistic biasing towards or against using tone contrasts linguistically must be seen as a "free" neutral by-product of these genes. It is, thus, plain non-sense to talk about the "superiority" of some linguistic systems based on the putative natural selection involved in these haplogroups.

Concerning the specific mechanisms linking the presence of *ASPM-D* and *MCPH-D* in a certain proportion of speakers in a populations and the usage of tone contrasts in the language(s) spoken by than population, at this stage, only speculations can be made. Until the directionality of this bias is clarified (towards non-tonality or against tonality) and its strength assessed, nothing more can be said than that it could involve anything from specifically linguistic to general neuro-cognitive processes, including fine temporal resolution or phonological working memory (see also Chapter 5).

4.10. The geographical patterning of linguistic diversity

That linguistic diversity is geographically patterned hardly needs arguing, but the mechanisms are highly debated, generally including a combination of historical and geo-ecological factors. The approach sketched in this Chapter offers the hope of disentangling the two main processes of inheritance from a common ancestor and areal transfer. This is done by controlling for geographical proximity and historical linguistic closeness and trying to find those linguistic features more prone to diachronic conservatism from those more

prone to areal coherence and both from those too labile to carry any such relevant information. Then, these features can be used as preliminary tools in the study of not yet systematically studied groups of languages, in the hope of identifying linguistic areas and genetic groupings, but this needs much more research and better sampling.

Another very important consequence is highlighted by the study of linguistic variation at different geographic scales. For example, it was unexpected to find that some linguistic features seem to reach their maximum dissimilarity at around 7500km and then to seem to converge at larger scales. If confirmed, this phenomenon might be taken as an index for successive large-scale linguistic and demographic processes, akin to Nichols' hypotheses (Nichols, 1992), or might be due to geo-ecological properties of the Old World, or some form or neighbor-inhibition.

4.11. Conclusions and future work

These competing explanatory hypotheses need specifically-targeted studies for their attempted falsification. They can be broadly classified as:

- *Better statistical analyses*, including more refined sampling and linguistic coding, enriching and refining the methodology presented in this chapter. Such an approach is needed in order to ascertain the impact of ecological boundaries, the strength of correlations, the nature of the putative demographic effects and the relative contribution of vertical and horizontal processes in shaping linguistic diversity. There are two main requirements:
  - *Better sampling*: especially the number and distribution of populations, but also more refined linguistic features and better coverage of the genetic data;
  - *Better techniques*: while the methods presented in this chapter are promising, more work is required to transfer spatial statistical methods into linguistics without violating any of their assumptions. But, no matter how costly such a process might be, its effects for quantifying the linguistic diversity and testing hypotheses are invaluable;
- *Studies of non-tonal L_1 speakers acquiring a tonal L_2*: is there a correlation between the individual abilities to process tone distinctions and the possession of these two

derived haplogroups? If such a correlation is found, then we will have a firm explanatory ground for understanding this type of bias as well as an invaluable genetic/molecular tool for dissecting the linguistic capacity, probably better suited than the catastrophic mutations of *FOXP2*;

- *L_1 (acquisition) studies*: is there is any difference in the processing/acquisition of sequential/parallel structures between carriers and non-carriers of the derived haplogroups? This is a related, but not necessarily identical question to the previous, given that the biasing effects of *ASPM-D* and *MCPH-D* could be active only during the critical period[264], or only during adult language acquisition[265], or both;

- *Computer simulation and mathematical modeling*: These will represent, together with the experimental studies, a very important direction of research. They will be used to answer questions like: What are the conditions which would allow such a bias to be manifested in language change, given the complexities of the other determining factors (language internal, multilingual situations, etc.)? Are the biases inferred from such modeling compatible with suggestions from biology? What are the evolutionary dynamics allowed by such models? What scenarios of language evolution do they favor? What is the probability of obtaining such correlations just by chance, given appropriate models? What is the distribution of linguistic features' stability through time? Probably, the most useful will be agent-based computer simulations, whereby the entire process, staring with the individual and ending with a pattern of genetic diversity, can be dissected and studied, but also large-scale, population-based models, could prove useful in specific situations.

It can be concluded, thus, that the methods developed in this chapter will prove useful in the study of genetic and linguistic diversity and the interactions between them, that it is plausible that there are differences in temporal stability between linguistic features, and that tone, as a manifestation of sequential/parallel linguistic processes, could be causally connected to the frequency of *ASPM-D* and *MCPH-D*.

264Providing a child language acquisition directed explanation, in the vein of Simon Kirby & Jim Hurford's original Iterated Language Model (Kirby & Hurford, 2001; Smith, 2004; Smith, Kirby & Brighton, 2003).

265Providing a language-shift explanation, more akin to Ostler (2005).

5. Non-spurious correlations and language evolution

This chapter will analyze in detail the generic theory of non-spurious correlations between genetic and linguistic diversities, as well as its particular case of *ASPM-D*, *MCPH-D* and *Tone*. It will be argued that such phenomena are one of the keys to understanding language evolution in the context of human evolution and that genetic and linguistic diversities are an essential aspect of our species. It will close with some concluding remarks, putting the entire thesis in perspective.

5.1. The theory of non-spurious correlations between genetic and linguistic diversities

Since the seminal work of Cavalli-Sforza and colleagues (Ammerman & Cavalli-Sforza, 1984; Cavalli-Sforza, Menozzi, & Piazza, 1994; Cavalli-Sforza, Piazza, Menozzi & Mountain, 1988; Cavalli-Sforza, Piazza, Menozzi & Mountain, 1989; and not only them – Chapter 3), it became generally accepted that there might exist correlations between the genetic and linguistic diversities on various scales. These correlations are to be mainly attributed to *demographic events* shaping the two types of diversity in roughly the same way:

> What explanations can one offer for this important correlation [between linguistic and genetic trees]? The *major explanation is the history of populations*. The correlation is certainly *not due to the effect of genes on languages*; if anything, it is likely that there is a reverse influence in that linguistic barriers may strengthen the genetic isolation between groups speaking different languages. [...] It is crystal clear that all normal human beings have essentially the same skills in learning languages, and the native tongue of an individual is essentially determined by the social environment in which the cultural development of that individual has taken place. [...] The explanation of the parallelism between genetic and linguistic trees is to be sought in the *common effect of factors determining differentiation both at the genetic and at the linguistic level*. The most important factors are *events determining the separation of two groups*. [...] It is reasonable to assume that *both the genetic and the linguistic divergence thus determined will increase with time since separation* (Cavalli-Sforza, Menozzi & Piazza, 1994:101, *italics* mine),

while a secondary component usually mentioned is represented by the causal feedback from linguistic diversity into genetic diversity through a process of *linguistic assortative mating*. Therefore, most of these correlations are *spurious*, in the sense that they are, in fact, explained by correlations with a third variable (de Vaus, 2002:316-318). A graphical

representation of this *language-genes standard model* (LGSM) is (Figure 69):

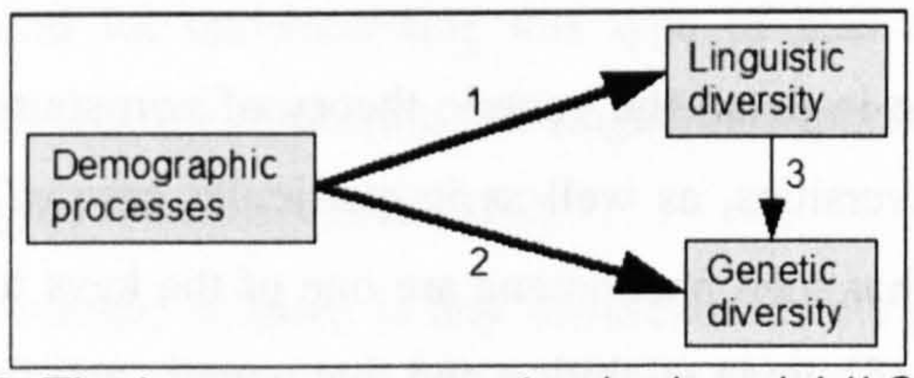

Figure 69: The language-genes standard model (*LSGM*).
[as advocated by Cavalli-Sforza, Menozzi & Piazza (1994)].

It must be highlighted that this *successive splitting view*, both at the linguistic and demographic (genetic) levels, can lead only to the standard way of considering correlations between genetic and linguistic diversity, namely inter-population variation in allele frequencies versus linguistic groups (dialects, languages, language families, etc.), which is reinforced by the secondary mechanism of linguistic assortative mating, implicitly assuming boundaries between such linguistic entities. This is why the field is more or less formally known as "language-genes correlations", as the only conceivable relationships allowed by this paradigm are between linguistic groups ("languages") and allele frequencies ("genes").

But the fundamental source of this paradigm, its most profound assumption which logically entails all its other assumptions, methods and interpretations, is that the *capacity for language*[266] *is uniform across the entire human species*. This assumption hardly needs any detailed discussion, being one of the first and most prevalent pieces of information acquired by undergraduate linguists during their training, and permeating all our work. Nevertheless, this form of uniformitarianism, which I will denote as *linguistic capacity uniformitarianism* (LCU), is currently in need of reevaluation.

First, LCU is different from the claim, with which it is sometimes mistakenly confused, that all human languages are equal (which I will denote as *linguistic communicative power uniformitarianism*, LCPU). To cite a leading linguistics introductory textbook (O'Grady, Dobrovolsky & Katamba, 1997):

> All languages and all varieties of a particular language have grammars that enable their speakers to express any proposition that the human mind can produce. In

266Loosely understood as what makes humans able to natively acquire and use a language. This must be regarded as a *generic* concept, not connected to specific (and reductionist) claims, like, for example, Chomsky's (1965) "Universal Grammar", or Hauser, Chomsky & Fitch's (2002) "Recursion" - see also Parker (2006b) for critiques.

> terms of this all-important criterion, then, *all varieties of language are absolutely equal as instruments of communication and thought* (p. 6, *italics* mine),

and it seems reasonable to accept its essence as valid. However, there are arguments against too strong an interpretation of this principle (see, for example, Gill, 1994, 2004).

Second, the content of the LCU must be amended in one very important respect: there are many types of pathology affecting language acquisition and some of them have a genetic component (Section 3.1). Therefore, the universality of the LCU must be restricted to "*normality*", as, for example, in the fragment from Cavalli-Sforza, Menozzi & Piazza (1994:101) above. But the problem of defining this "normality" creeps in, aggravated by the conception of pathology as the tail end of the distribution as opposed to a clear-cut distinct category (Section 3.1.6). Therefore, it is at best unclear what the *real* content of this assertion (LCU) is, given that the distribution of the capacity for language does not fall in two distinct classes (normal and pathologic). Moreover, the genetic components involved in it point to a complex mosaic, each with its own statistical distribution (Section 3.1). Therefore, it must be concluded that LCU is, at the worst, a vacuous claim, and, at best, a discretized abstraction of a very complex statistical reality.

The difficulty of accepting the existence of inter-individual differences concerning the linguistic capacity is disturbing, given the enormous amount of relevant data (Section 3.1), but it seems that the widespread recognition of this inter-individual diversity is inevitable. Probably, this resistance is due to a deeply entrenched misunderstanding of human diversity and a misplaced and exaggerated counter-reaction to past and present discrimination, very much akin to the political correctness pressures on human evolution (Annex 2). It must be highlighted that a recognition of inter-individual differences might allow for less discrimination than a blind and absolutist, almost religious, claim to universal and indiscriminate uniformity; in the same vein, agenda-motivated arguments and accusations have no place in a scientific debate (Annex 2). Moreover, from a biological point of view, the existence of inter-individual variation in the linguistic capacity is natural, given the overwhelmingly important role played by variation in most biological accounts, and, given the data presented in Chapter 3, some of this variation is to be attributed to genetic variation.

The paradigm shift allowed by such a *variationist* point of view on the human capacity for

language is potentially far reaching and important. One such impact, concerning the correlations between genetic and linguistic diversities, is represented by the possibility of *non-spurious correlations* besides the familiar accidental ones.

5.1.1. The (fictional) example of [r] and [ɹ]

In the standard paradigm, the correlation between the frequency of a certain allele, *A*, across various populations and the linguistic entities spoken by those populations, is due to demographic effects and/or linguistic assortative mating, thus, being mostly spurious, and it is also *random* as to the nature of allele's *A* effect. More explicitly, it is assumed that the effects of this allele do not change the probability of its correlation with languages, except through a demographic intermediate. In this sense, the correlation is *accidental*.

But, in the new paradigm, it can be envisaged that this allele, *A*, has a *biasing effect*, affecting the relative probabilities of a set of linguistic variables (features). This biasing effect is present at the level of the individual and, depending on various factors, can become manifest or not. For example, let us suppose that this allele, *A*, affects the articulatory easiness of producing the alveolar trill sound (IPA [r])[267], in the sense that its carriers have a higher probability than non-carriers, $p_c > p_{nc}$, of not acquiring the capacity to produce this sound. It must be pointed out that this example is not entirely fictional, as there seems to be a genetic component in the inability of certain speakers to articulate [r]:

> The results of these analyses suggest that *articulation of the phoneme /r/ is largely the result of genetic factors*, whereas environmental factors play a greater role in the articulation of the phonemes /l/, /w/, and /j/ (Stromswold, 2001:673, *italics* mine)[268].

The factors modulating p_c relative to p_{nc} can include exposure, specialized training, social conformist pressure, disease and many other environmental (and explicitly cultural) effects. This type of variation is not pathologic and represents, thus, a normal polymorphism in human populations.

Now, zooming out from the individual carrying the allele *A* to the containing speech

---

267This simplistic, one gene-one phene model is used just for illustration purposes. It is highly probable that, in reality, there are many genes with small effects influencing the articulation of [r] and, also, that there are phenocopies involved (West-Eberhard, 2003).

268Thanks to Mits Ota for comments.

community, a series of scenarios can be envisaged. The simplest one involves a low frequency of allele *A* in the population. In this case, its carriers will manifest the phenotype with probability p_c and fail to acquire the articulation of [r]. Depending on the language, *L*, spoken by the population, there are five possible cases:

i. if *L* does not make use of [r], then the allele *A* remains invisible;

ii. if *L* does use [r] (like Spanish, Romanian, Italian or Russian, for example), then *A* becomes visible with probability p_c, by forcing the manifest carriers to systematically replace [r] by a best approximation, including the alveolar approximant [ɹ][269];

iii. if *L* uses [r] and [ɹ] as allophones (like Yoruba), then allele *A* remains hidden in most everyday situations;

iv. if *L* phonemically contrasts [r] and [ɹ] (like Armenian and Albanian), then, presumably, the manifest carriers have a disadvantage by introducing supplementary homophony;

v. if L uses only [ɹ] (like Swedish), then the allele *A* also remains invisible.

If, for some reasons, including random genetic drift or natural selection on other phenotypic effects, the frequency of *A* in the population increases, then, assuming p_c constant, the frequency of manifest carriers will increase. For cases (i) and (v), this will have no effects on *L*, but for the other cases, this could possibly determine a language change, whereby, a type (ii) language will become a type (iii) language through the systematic introduction of the allophony [r] – [ɹ] and, presumably, a type (iv) language will also become a type (iii) language through collapsing the two phonemes [r] and [ɹ] into one. If we imagine a further increase in allele *A* frequency, tending towards fixation, type (iii) languages will possibly converge into type (v), as [r] drops out of use. Thus, an abstract depiction of this process, represented as the probability of types (i)-(v) languages function of the frequency of *A*, is given in Figure 70:

269I am myself a native speaker of Romanian and unable to articulate [r], systematically replacing it by [ɹ].

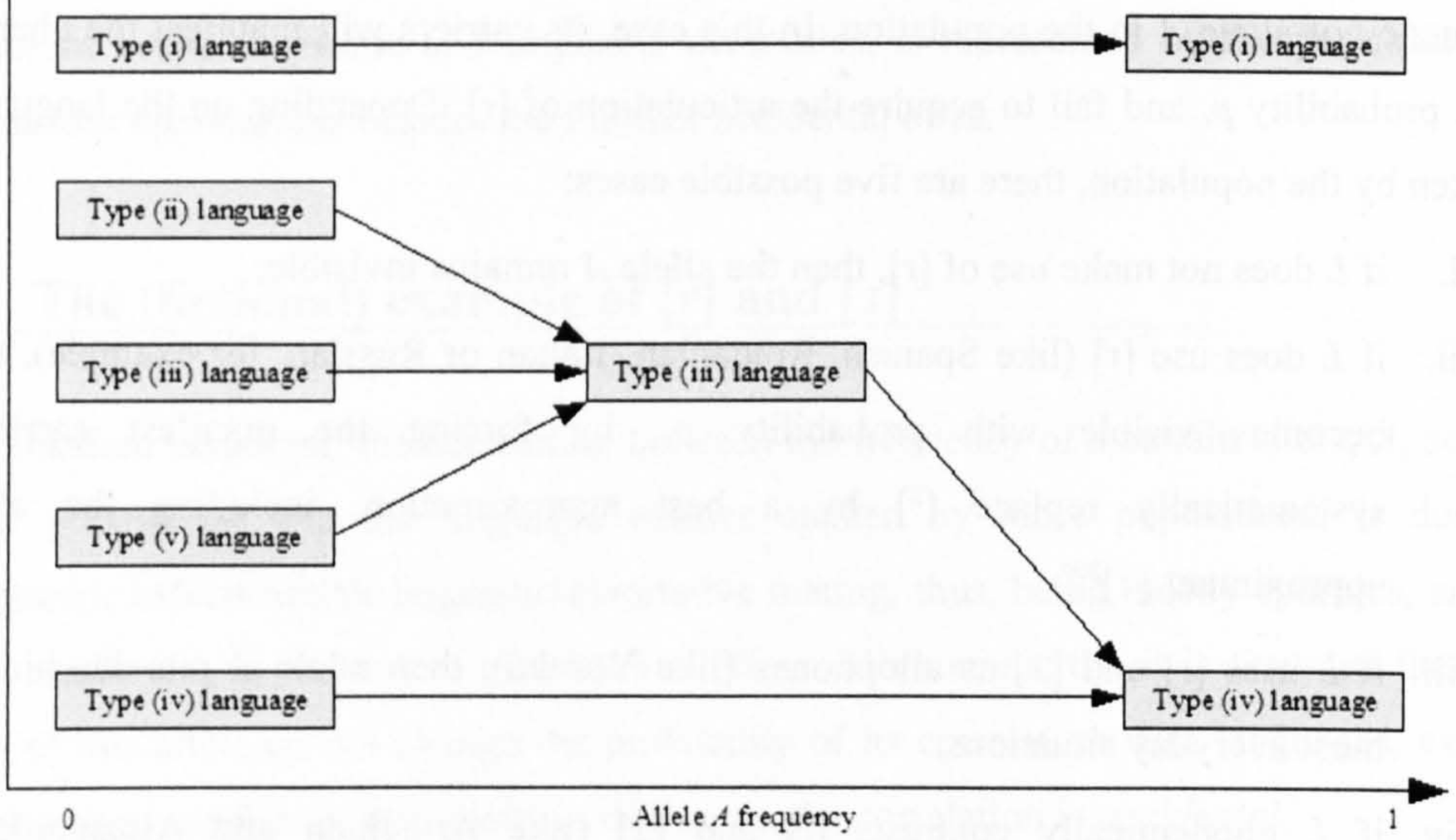

Figure 70: Language types distribution with increasing allele *A* frequency.
Type (i): no usage of [r]; type (ii): [r] only; type (iii): [r] and [ɹ] allophones; type (iv): phonemic contrasting [r] and [ɹ] and type (v): [ɹ] only.

Thus, the five types of languages possible in the absence of allele *A* collapse into only two stable types as *A* reaches fixation. It must be highlighted again that this represents an exaggerated and simplified model, used for illustrative purposes only. Future modeling work must take into account the fact that the bias induced by *A* is potentially very small and that the effects on *L* are not linear with *A*'s frequency in the population, along with many other factors. Nevertheless, that main idea is that *a genetically motivated small bias, manifested as inter-individual diversity, can lead to language change*, the genetic factor acting as a *constraint on linguistic transmission*.

The dynamics of this system can be further complicated by integrating a feedback selective pressure from *L* onto *A*, whereby manifest carriers unable to articulate [r] suffer a biological fitness increase or decrease relative to the others. For this to be possible, they must be visible, which involves cases (iv), (ii), and possibly (iii), while the plausible selective mechanisms can include sexual selection or social norms. If so, various scenarios are possible, but further mathematical and computational modeling is required to asses their probabilities. Nevertheless, a plausible consequence is that positive selection of manifest carriers will lead to case (v) languages, while negative selection will lead to higher frequency of allele *A* in populations speaking type (i) or (v) languages. But if we limit ourselves, for the

moment, to selectively neutral linguistic effects, another interesting case is represented by genetic biasing of two reciprocally exclusive equivalent values of a linguistic feature.

5.1.2. The case of *tone*

Tone represents a very complex linguistic phenomenon: acoustically, it is based mainly on the *fundamental frequency* (F0), while the vowels and consonants are based primarily on the *spectral properties* of speech (Fry, 1979; Crystal, 1975), which can be distinguished in auditory perception, and which forms the basis of the linguistic distinction between tone and segmental sounds (Cutler, Dahan & van Donselaar, 1997). *Pitch* is organized into *tone phonemes*, which represent an important part of the phonological form of morphemes. Tone phonemes may be *levels* (register tone) (e.g., High, Medium, Low) or *contours* (e.g., Rise, Fall) (Maddieson, 2005; O'Grady, Dobrovolsky & Katamba, 1997:44-45), which, in many tone languages, are phonologically composite (e.g. R = LH). In standard tone languages, every syllable has a tone phoneme in addition to segmental phonemes and in many tone languages, some morphemes consist only of a tone phoneme, while in many others, some grammatical functions are signaled by tone changes. Examples of *lexical* and *grammatical* tone contrasts (Maddieson, 2005; Dobrovolsky & Katamba, 1997:44-45; D.R. Ladd, *pc*) are:

Language	***Form***	***Gloss***	***Type***
Mandarin Chinese [cmn]	ma^{H}	mother	Lexical tone
	ma^{MH}	hemp	
	ma^{MLH}	horse	
	ma^{HL}	scold	
Sarcee/Sarsi [srs]	$mi^{H}ɬ$	moth	Lexical tone
	$mi^{M}ɬ$	snare	
	$mi^{L}ɬ$	sleep	
Yoruba [yor]	$i^{M}gba^{H}$	calabash	Lexical tone
	$i^{L}gba^{H}$	a tree	
	$i^{M}gba^{M}$	200	
	$i^{L}gba^{L}$	time	
Bini/Edo [bin]	$i^{L}ma^{L}$	I show	Grammatical tone: tense
	$i^{H}ma^{L}$	I am showing	
	$i^{L}ma^{H}$	I showed	

Table 41: Examples of lexical and grammatical tones.

Superscripts L, H, and M refer to tone.

Geographically, languages using tone contrasts have a skewed distribution, being very well represented in sub-Saharan Africa, East and South-East Asia and Central America/Caribbean/Amazonia (Maddieson, 2005). Typologically, between canonical tone languages (Yoruba, Thai, Vietnamese, Dinka, Mixtec, etc.) and canonical non-tonal languages (English, Arabic, Tamil, Indonesian, Hungarian, etc.) there are a number of intermediate cases, often called *pitch accent* languages, which often involve pitch distinctions that are limited to a specific syllable in a word (e.g., Swedish/Norwegian, Lithuanian, South Slavic, Basque, Japanese, etc.) (Maddieson, 2005; Section 5.2.4). Historically, languages can become tonal through internal processes of sound change and phonological reanalysis (e.g., Swedish/Norwegian, Chinese), those that are already tonal can acquire new tonal contrasts through phonological reanalysis (e.g. voicing distinctions in Chinese) or can lose tone distinctions, especially in contact situations (e.g., Swedish in Finland and Swahili as a trade language).

Simplifying, tone can be used to produce a binary classification of languages into a class using tone contrasts and another one, composed of languages not using tone contrasts (Chapter 4; Maddieson, 2005). Now, if we consider this linguistic variable (feature) as having two possible values, 1 for the first class and 0 for the second, these two values are absolutely neutral from a linguistic point of view, meaning that a language using tone distinctions is perfectly equivalent (expressively) to a language not using tone distinctions. This concept of linguistic neutrality is very important, as it highlights the fact that the choice of specific values for certain linguistic features are, on functional grounds, equally probable, as opposed to some other features, which could have non-neutral values[270]. An example of this last type is represented by the linguistic feature *center embedding*, which represents the depth of phrasal center embedding: as is well known (e.g., Kirby, 1999), functional considerations strongly constrain its values.

The default assumption is that the linguistic capacity is not biased for or against any value of tone, but, let us assume that allele *A*, when present in an individual, biases, for example, the perceptual strategies for separating F0 from spectral cues, increasing the probability of interpreting pitch differences as tone distinctions (this is but one possible mechanism; see below). Therefore, at the population level, when the frequency of *A* is high enough, the

270But such a decision, once made, will impact on other aspects of the language.

language *L* could be changed towards introducing more tonal distinctions.

5.1.3. From individual genetic biases to language change

It must be noted that the causal gap between individuals carrying the allele *A* and the eventual language change in the direction of the bias induced by *A* is bridged by *cultural transmission*, understood as a complex, active and noisy transmission mechanism, potentially amplifying (or masking) these biases (e.g., Smith, 2003; Smith, 2004; Dowman, Kirby & Griffith, 2006). Therefore, the causal connection between the individual genotypes and the community's language is mediated by an inter-generational process of *iterated learning* by (potentially) *biased learners* from (potentially) *biased producers*. This process requires a certain structuring of the population (including allele frequency) and iterated cultural transmission (bias attenuation or amplification), demanding a certain number of generations to full manifestation. A supplementary, but very important complication is added by the *dynamics* of allele *A*'s frequency. This process, an extension of the classical *Iterated Learning Model* (Kirby, 2001; Hurford, 2002), and denoted *Genetically Biased Structured Iterated Learning* (GBS_{IL}), can be captured in the following diagram (Figure 71).

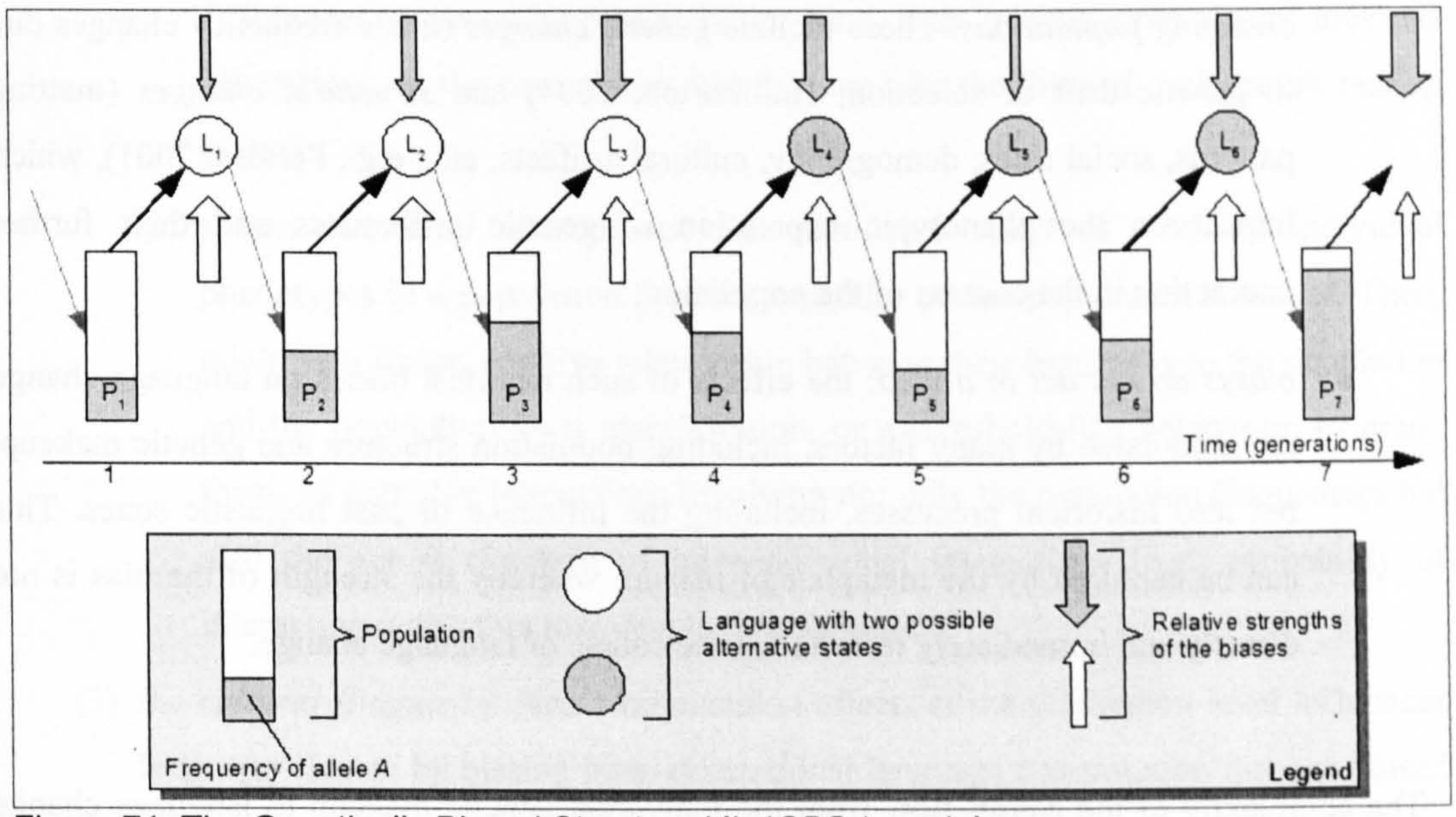

Figure 71: The Genetically Biased Structured IL (GBS_{IL}) model.

Language transmission through repeated learning across generations is amended by genetic biases of individuals in structured populations.

In this simplified representation of the GBS_{IL}, time is discretized in generations, P_1, ..., P_n represent the population at times 1, ..., n, and L_1, ..., L_n represent the language spoken by the population. The main feature of an iterated learning model of language is that each generation's language is the product of the previous generation's language (*expression*) through active learning (*induction*), and the extension brought by GBS_{IL} is that *the expression, induction, or both, are influenced by genetic biases within a structured population*. It must be pointed out that the problematic of the influence of genetic (i.e., pre-existing) biases on the trajectory of language change through ILM is as old as the ILM itself, and very thorough explorations are, for example, Smith (2003, 2004), but the emphasis specific to GBS_{IL} is on *inter-individual variation in the expression of this bias* (as opposed to uniform expression in a population). Figure 71 above tries to depict a series of effects of GBS_{IL}:

- *inter-individual variation*: as opposed to most models of language change (ILM included), the accent is on the *non-identity of individuals as linguistic agents*. They are different genetically, and the phenotypic manifestation of these differences is *mediated by complex factors* (both genetic and environmental; Chapter 3);

- *population dynamic structure*: it is not enough to recognize the importance of inter-individual differences, but these differences are embedded into *complex, structured, changing populations*. These include *genetic changes* (allele frequency changes due to genetic drift or selection; Halliburton, 2004) and *structural changes* (matting patterns, social rules, demography, cultural artifacts, etc.; e.g., Ferraro, 2001), which impact on the phenotypic expression of genetic differences and their further interaction in the context of the population;

- *biases do not act in a void*: the effects of such manifest biases on language change are modulated by many factors, including population structure and genetic makeup, but also historical processes, including the influence of past linguistic states. This can be captured by the metaphor of *inertia*, whereby the strength of the bias is not directly and immediately reflected in the course of language change.

The complexity of the causal flow from the individual genetic makeup to language change can be split into several stages:

(1) *the individual level*: here, the causal gap between genetic makeup and manifest

phenotype must be bridged:

(1.1) *the genetic makeup*: the individual must possess the appropriate alleles, either through inheritance or *de novo* mutation. This can be represented by changes in regulatory or structural genes, by patterns of DNA methylation, etc., but probably involving many genes with small effects (Chapter 3);

(1.2) *the phenotypic penetrance*: a plethora of factors impact on the phenotypic manifestation of genetic differences (for good reviews, see West-Eberhard, 2003, Gerhart & Kirschner, 1997, or Lewin, 2004; Chapter 3) and include dominance/recessiveness, pleiotropism, epistasis, environmental effects (masking, amplification) and organismal plasticity (the "extended phenotype"-like and cultural effects are discussed below);

(2) *the population level*: population-level phenomena impact both on the individual phenotypic manifestation of genetic differences and on the population-scale dynamics of such manifest individual phenotypes:

(2.1) *"extended phenotype"-type effects*: population-level phenomena can impact on the individual phenotypic manifestation of genetic differences through extended phenotype (Dawkins, 1982) or niche construction (Odling-Smee, Laland & Feldman, 2003) type of effects, whereby the current penetrance of genetic differences is biased by past activities (and, thus, genetic structure) of previous generations. In the case of humans, this can take the form of social rules, mating systems, etc.;

(2.2) *inter-individual interactions*: the cumulated effect of manifest individual phenotypes in a population depends crucially on the population structure. There might be a linear, additive relationship between their frequency in the population and the population-level manifestation, or a threshold-like behaviour, or many forms of complex interactions involving not only the population frequencies but also the actual structure of inter-individual interactions (e.g., probability of interacting with other manifest individuals, etc.);

(3) *the cultural (linguistic) level*: the manifest effects at the population level influences language change by biasing inter-generational language transmission through biased expression, induction, or both, in the context of the pre-existing linguistic structures. This step involves complex interactions between the manifest (population-level) biases and the previous historical processes shaping the population's language(s) and

potentially results in a specific linguistic change trajectory.

These processes are represented in Figure 72 (the inter-population influences will be discussed later).

It must be pointed out that, due to this complexity of interaction, very small individual biases can be amplified by socio-cultural (iterated) processes, leading to visible effects (as suggested, for example, by Dowman, Kirby & Griffith, 2006), or, *au contraire*, relatively large biases can be damped. For us, the most interesting is the first, amplificatory, case, whereby the social and cultural dynamics on an explicitly temporal dimension can transform the small individual biases into visible linguistic effects. Nevertheless, given these complexities, further mathematical and computational modeling is needed for a detailed understanding of this type of processes and their plausible effects.

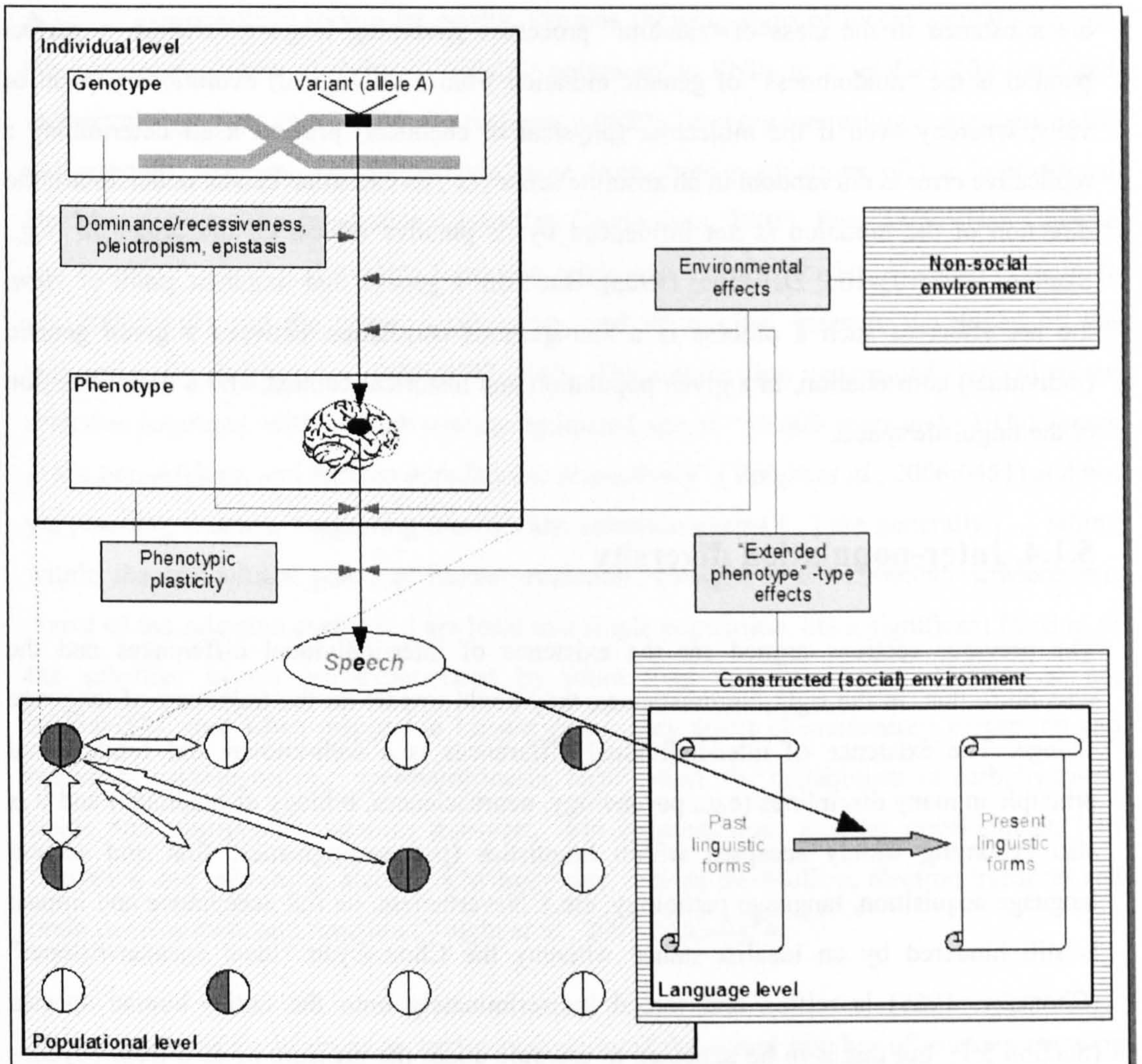

Figure 72: Schematic representation of the complex modulation of the causal links from an individual's genome to language change.

Many simplifying assumptions have been made, which are not necessarily true in the general case (e.g., the phenotypic effect can be purely organic – articulatory or acoustic, etc.).

Another important observation is that, from a purely linguistic point of view, a language change determined by such a process is *not different in any respect* from other internally motivated language changes. Therefore, from a linguistic point of view, these mechanisms

are subsumed to the class of "random" processes governing language change; a perfect parallel is the "randomness" of genetic mutation from a (biological) evolutionary point of view, whereby even if the molecular (physical or chemical) process itself determining a replicative error is not random in an absolute sense (i.e., its causation can be understood), the direction of the mutation is not influenced by its putative effects on the organism (e.g., Skelton, 1993:103-106; Dawkins, 1990a). But from a genetic and linguistic point of view, the net effect of such a process is a non-spurious correlation between a given genetic (individual) constellation, in a given population and historical context, and a defined region of the linguistic space.

5.1.4. Inter-population diversity

The previous sections argued for the existence of inter-individual differences and the possibility that, in the right circumstances, they could impact on the trajectory of language change. The existence of inter-individual differences is a well-known and fundamental principle in many disciplines (e.g., psychology, neurosciences, biology or medicine) and it is also becoming widely accepted within linguistics (psycho-linguistics, first and second language acquisition, language pathology, etc.). Nevertheless, its full acceptance and impact is still hindered by an idealist stance whereby the Chomskyan "ideal speaker-listener" (Chomsky, 1965) is reified and forced indiscriminately onto the entire human species (Section 5.1), but this is to be expected to crumble under the pressure of data from psycho-linguistics, socio-linguistics, neuro-cognitive sciences and many others.

As argued at length in Section 2.2.8, there is enough genetic structure in the living human populations to allow, for example, individual identification or targeted medicine. But it must be highlighted again that this structure does not, in any way, support any racist claims of any form (Chapter 2, Annex 2; below), but it must also be pointed out that purely political or moral agendas cannot be used to deny its existence (Annex 2). It is known that the set of genetic inter-population variation is vast (e.g., Jobling, Hurles, Tyler-Smith, 2004), ranging from neutral markers (SNPs, STRs, etc.), to intensely naturally selected (skin color, sickle-cell anemia or lactose-tolerance). While it is possible that some neutral polymorphisms might have linguistic effects, this is in principle very improbable, and, therefore, we will focus, in the following, on non-neutral genes.

A recent genome-wide survey for genetic variants showing signs of recent natural selection (Voight *et al.*, 2006), studying "~800,000 polymorphic SNPs in a total of 209 unrelated individuals" (Voight *et al.*, 2006: Corrections, p.0659), has very interesting consequences for our research program. The individuals came from three populations of the International HapMap project (The International HapMap Consortium, 2005): East Asian (Han Chinese and Japanese), north and western European origin and Yoruba (Ibadan, Nigeria) (Voight *et al.*, 2006:0447) and the genome-wide scan used an original measure developed by the authors (*iHS*, Voight *et al.*, 2006:0447-0449). The authors find widespread signs of recent selective pressures, with a rough average estimated age of "≈6,600 years and ≈10,800 years in the non-African, and African populations, respectively" (Voight *et al.*, 2006:0451) and not yet reaching fixation, suggesting that "[...]the selection events [...] are generally [...] falling within the agricultural phase of human evolution" (Voight *et al.*, 2006:0451). Moreover, "most of the selective events [...] are local to a single population, but a significant fraction of the selective events are experienced by more than one population" (Voight *et al.*, 2006:0452), and, when mapped to known genes, they cover chemosensory perception and olfaction, gametogenesis, spermatogenesis, fertilization, the metabolism of carbohydrates, lipids and phosphates, vitamin transport, skin pigmentation, skeletal development, hair formation and patterning, alcohol dehydrogenase, lactose metabolism, electron transport and brain development and function (Voight *et al.*, 2006:0453-0454).

This study must be taken as an under-evaluation of the actual number of markers showing signals of natural selection, given that the methodology used by the authors "[...] is aimed at finding loci where there is strong, very recent, selection in favor of alleles that have not yet reached fixation" (Voight *et al.*, 2006:0446), favoring, thus, very recent processes, and that the sampling is very limited. Nevertheless, it clearly shows that the genetic structure of human populations is not restricted to neutral loci, but, as expected on evolutionary grounds, also includes naturally selected genes, pointing to both globally relevant selective pressures as well as to more local/regional ones.

But while it is highly improbable that reports of natural selection for skin color or lactose tolerance will generate any uneasiness, this is not to be expected for genes concerning the brain. Nevertheless, while selected genes, showing inter-population patterning, but not involved in brain morphology or activity might still prove to have linguistic effects (by, for

example, influencing the shape and dynamics of the articulatory organs), it is this latter class of genes which promises to be most interesting for us. Therefore, we have to put aside any misplaced political or moral arguments and further explore this path.

5.1.5. Genes showing signs of natural selection, inter-population patterning and involvement in brain development and/or functioning

From a purely evolutionary point of view, given that the human brain and the processes it supports represents one of the most dynamic aspects of human evolution, it is to be expected that natural selection on its different components has not brusquely and magically ceased at a conventional (and convenient) moment in prehistory. However, finding genes showing signs of natural selection and somehow involved in brain development and/or functioning is but the first step towards establishing their actual function(s) which are under natural selection, and the understanding of the nature of the selective pressures involved.

The Voight *et al.* (2006:0454; see above) genomewise study did find signals of very recent natural selection on genes involved in microcephaly (*CDK5RAP2*, OMIM 608201, *MCPH3*, selected in their Yoruba sample, and *CENPJ*, OMIM 609279, *MCPH6*, selected in their European and East Asian samples), primary inhibitory neurotransmitter *GABA* (*GABRA4*, OMIM 137141, selected in their Yoruba sample and possibly involved in autism), a susceptibility to Alzheimer's disease gene (*PSEN1*, OMIM 104311, selected in their Yoruba sample), a gene involved in Ca^{2+} binding and synaptic functioning (*SYT1*, OMIM 185605, also selected in their Yoruba sample), the serotonin transporter gene (*SLC6A4*, OMIM 182138, selected in their European and East Asian samples, and which actively removes serotonin from the synaptic space) and the dystrophin binding gene (*SNTG1*, OMIM 608714, in all three samples), showing the pervasiveness of natural selection on brain development and/or functioning, presumably due to cognitive, emotional or other psychological effects. Most such selection signals seem to be local (only the last appears in all three samples) and they involve a wide range of aspects regarding the brain. Possibly, some of these might also have linguistic effects of the type advocated in the previous sections, but, for the moment, there is not enough data to test this hypothesis.

Also, Wang *et al.* (2006), by applying a different (probabilistic) method to SNPs, identified regions with a high probability of recent positive selection (10-40 ky) (Wang *et al.*, 2006:137), and mapped them to known genes (where possible). Out of the 112 genes indicating selection in all tested populations (European Americans, African Americans and Chinese from the Los Angeles area), they found (Wang *et al.*, 2006:139) 7% involved in reproduction, 10% in host-pathogen interactions, 13% in cell cycle, 15% in protein metabolism, 17% in neuronal function (e.g., serotonin transporter *SLC6A4* – also found by Voight *et al.*, 2006; glutamate and glycine receptors *GRM3* [OMIM 601115], *GRM1* [OMIM 604473] and *GLRA2* [OMIM 305990]; olfactory receptors *OR4C13* and *OR2B6*; synapse-associated proteins like *RAPSN* [OMIM 601592] and others, like *ASPM* or *RNT1*) and 21% in DNA metabolism.

But the best-known such genes, are, without doubt, *ASPM* and *Microcephalin* (Evans *et al.*, 2005; Mekel-Bobrov *et al.*, 2005; Section 4.2). As discussed in Section 4.2, they show signs of natural selection, inter-population structure, and are involved in high-functioning microcephalias (Evans *et al.*, 2005; Mekel-Bobrov *et al.*, 2005; Gilbert, Dobyns & Lahn, 2005; Mekel-Bobrov *et al.*, 2006). Their effects on brain size and development are inferred primarily from the fact that their deleterious mutations are associated with primary microcephaly:

> [...] mutations in this gene [*Microcephalin*] cause primary microcephaly [MCPH; Online Mendelian Inheritance in Man (OMIM) accession 251200] [...]. MCPH is defined clinically as severe reductions in brain size coupled with mental retardation, but remarkably, an overall retention of normal brain structure and a lack of overt abnormalities outside of the nervous system (Evans *et al.*, 2005:1717)

and

> [h]omozygous null mutations of *ASPM* cause primary microcephaly, a condition characterized by severely reduced brain size with otherwise normal neuroarchitecture (Mekel-Bobrov *et al.*, 2005:1720).

More specific suggestions are that "*Microcephalin* is suggested to control the proliferation and/or differentiation of neuroblasts during neurogenesis" (Evans *et al.*, 2005:1717) and "[...] *ASPM* may regulate neural stem cell proliferation and/or differentiation during brain development, possibly by mediating spindle assembly during cell division" (Mekel-Bobrov *et al.*, 2005:1720).

Thus, both *ASPM* and *MCPH* (for notations, see Section 4.2) seem involved in the control of brain growth and development, with potentially multifaceted phenotypic effects, including linguistic. Moreover, their derived haplogroups (*ASPM-D* and *MCPH-D*) are recent enough (5.8ky, 0.5-14.1ky 95% CI for *ASPM-D* and 37ky, 14-60ky 95% CI for *MCPH-D*) for their effects to be understandable within a modern cognition framework. Therefore, together with the 7 genes found by Voight *et al.* (2006), *ASPM* and *MCPH* represent prime candidates for non-spurious correlations with linguistic variables, but they alone have the major advantage of detailed enough world-wide sampling, making them ideal for a correlational study.

5.1.6. *ASPM, MCPH* and *Tone*

The main hypothesis, formulated shortly after the publication of Evans *et al.* (2005) and Mekel-Bobrov *et al.* (2005), during October 2005, jointly by prof. D.R. Ladd and me, concerns a causal relationship between the distribution of the derived haplogroups of *ASPM* and *MCPH* and the linguistic use of tone distinctions. The original source of this hypothesis was represented by the similarity of the distributional maps of the two haplogroups with the typological maps of tone, coupled with the brain effects of these genes, in the context of previous discussions concerning non-spurious correlations (me) and the special place of tone in the putative parallel/sequential linguistic system (Prof. D.R. Ladd). After initial checking of the plausibility of this hypothesis from the point of view of possible mechanisms, we decided to proceed with a statistical, correlational study. While vividly aware of the limitations of such a study, especially concerning causality, we decided that it represents the first logical step towards attempting the falsification of this hypothesis, eventually followed by more powerful (but also more resource-intensive) studies (Section 4.11).

The statistical techniques employed (Chapter 4) failed to reject this hypothesis and, more than that, allowed the formulation of a more specific version, whereby the frequencies of the two haplogroups in a population are related in a specific way to the use of tone distinctions (Section 4.9). This relationship holds even when geography and common linguistic descent have been controlled for, suggesting that this correlation is very important and real. The step from correlation to causation must await further targeted studies (Section 4.11), but seems to be a safe claim for the moment. If these findings will resist future, more powerful, tests, it would represent the first case of non-spurious correlation between genetic and linguistic

diversities. But even if these claims will prove false, the general theory of non-spurious correlations, the search for candidate genes between those showing signs of natural selection, inter-population diversity and brain involvement, coupled with the methodology developed in Chapter 4, still remain valid and in need of more systematic application.

5.1.7. Inter-population diversity revisited: why do we need it and what does it mean?

As shown in Section 3.1.2, estimating the relationship between the individual genetic makeup and phenotype requires inter-individual variation; the same principle applies to establishing non-spurious correlations between genetic and linguistic diversities, but at the population level. Therefore, inter-population variability, both from the genetic and linguistic points of view, is a fundamental requirement for such studies. The genetic variability refers to differences between the frequencies of the chosen *genetic variants* in the populations, while the linguistic variability refers to different values of the chosen *linguistic variables* (features) or to inclusion in different *linguistic entities* (languages, linguistic families, subfamilies, etc.) for the populations' languages.

While the linguistic inter-population variability requires no further discussion, the genetic variability, especially when, as argued in Section 5.1.4, refers to naturally selected loci, might raise certain problems. In order to exist, it requires differences in allele frequencies at these loci, which, for naturally selected loci, requires an explanation. It can be caused by differences in selective pressures due to physical, ecological or cultural differences (e.g., skin pigmentation, malaria resistance and lactose tolerance) or it can represent an ongoing increase in frequency towards fixation. This last mechanism seems especially prone to be hijacked by racist agendas, as proven by the *ASPM* and *MCPH* case (Section 4.2) and used to argue that certain "races" are "inferior" as certain "good genes" have not yet arrived. However, such arguments are simply irrelevant.

A very important point is that such inter-population differences, both genetic and linguistic, are *intrinsically dynamic*, as selective pressures continuously change and gene frequencies evolve (possibly) towards fixation, and languages change, become replaced and mixed. Nevertheless, it is exactly this dynamism which allows the manifestation and uncovering of

non-spurious correlations of the type discussed here. Thus, such an approach is essentially non-static, dynamic, based on continuous change, opposing fixed classifications and hierarchies. Therefore, accusations of hidden racism are, at best, contorted.

Another observation is that this does not offer in any way a hierarchical organization of languages, akin to racist "ladders" of humanity (Banton, 1998; Wolpoff & Caspari, 1997; Annex 2). Let there be a linguistic feature *f*, with values v_1, v_2, ..., v_n, non-accidentally correlating with a locus with two alleles: *A*, favored by natural selection, and *a*, in the process of being replaced. Let the values v_1, ..., v_i, be determined by allele *A*, while the remaining values v_{i+1}, ..., v_n, be determined by allele *a*. Then, a simple-minded claim would be to classify languages for which *f* uses values v_1, ..., v_i as "superior" to languages using values v_{i+1}, ..., v_n, but such an approach is unwarranted. Generally, it is to be expected that, from a linguistic point of view[271], values v_1, v_2, ..., v_n are equally functional, representing equally likely choices for feature *f*, and it is this neutral choice which is biased by *A* or *a*. Nevertheless, there exists the logical possibility that values v_1, ..., v_i are better on some functional criterion and that natural selection on *A* is (partially) due to this (as this will be argued to represent the main engine of language evolution), but even in this case, it is misleading to compare whole languages as opposed to specific linguistic features[272,273]. Therefore, hierarchically classifying languages on such a base is totally unwarranted and unscientific. Moreover, the theory of non-spurious correlations between genetic and linguistic diversities does not, in any sense, support the simplistic (and covertly racist) claims of the type "genes for Chinese".

And, finally, what does it mean that there are inter-population differences? First, in a metaphoric sense, this is one of the most important gifts made by Nature to man, as opposed to bleak uniformity. Second, it is not only an effect of spatial and cultural structure of the human species but also a direct result of different selective pressures, some of them possibly due to culture itself, through niche construction-type mechanisms (Odling-Smee, Laland, &

271 Which is the only one relevant in this discussion, as opposed to "aesthetic", "prescriptive", etc., no matter how aggressive their advocacy.

272 This is very much akin to trying to find absolute scales for comparing individual humans (the craze of the IQ seems, fortunately, to fade, under the pressure of data showing its unreliability, cultural loading and specificity), as opposed to specifically designed, domain-specific, measures.

273 As languages, representing functional linguistic systems, have innumerable ways of adjusting for such differences, very much like whole organisms can use phenotypic plasticity to cope with genetic (deleterious) effects (West-Eberhard, 2003).

Feldman, 2003).

5.1.8. The apparent paradox of too few non-spurious correlations

Given the large number of linguistic aspects (including pathologies) which have high heritabilities (Section 3.1.3), why is it that we don't see many more such non-spurious correlations? This question[274] is very interesting and its answer important for clarifying the theory of non-spurious correlations between genetic and linguistic diversities.

Let's consider a language-related condition with high heritability. Why, then, don't we find systematic inter-population differences in those linguistic aspects concerned by it, given that, with such a large estimate of heritability, slight relevant genetic differences would become manifest relatively easily? For example, stuttering (Section 3.1.3) is highly heritable ($h^2 \approx 0.70$), and is easy to imagine its effects on language change, but there is no non-spurious correlation involving it (nor am I willing to claim one!). These are the possible answers:

- first, enough individuals manifesting the relevant phenotype must be present so that there are discernible linguistic effects: in other words, in many such cases, given the deleterious nature of the phenotypes themselves or their association with deleterious phenotypes, their frequency is rather low;
- second, as discussed at length in Section 3.1.2, heritability estimates are not absolute and depend on environmental factors: therefore (see Section 5.1.3), many other factors must be considered, including the penetrance (heritability) in various contexts;
- third, as discussed in Section 5.1.7, inter-population differences in frequency must exist in order for non-spurious correlations to be found: there seem to be no data on inter-population differences in the frequency of most of the high-heritability factors discussed (Section 3.1.3), but this may be due to biased assessment. Nevertheless, if this uniformity of inter-population distribution is true, it might be the case that, as an extreme example, every language is in fact affected by stutterers, but because of this lack of variability, there is no (ethically acceptable) way of proving this non-spurious correlation(s);
- finally, it is highly possible that these high-heritability aspects of language are

274Thanks to Simon Kirby for asking this question.

influenced by many genes with small effects (Section 3.1): this would make the proof of a non-spurious correlation hard, as all (or the most important) such genes have to be identified and their frequency assessed in various populations.

It can be concluded, thus, that the lack of obvious non-spurious correlations involving high heritability linguistic aspects does not constitute a refutation of the general theory of non-spurious correlations between genetic and linguistic diversities.

5.1.9. What about the mechanisms?

The mechanisms bridging the causal gap between inter-individual genetic differences and biases potentially affecting language change belong to a diverse class. The simplest possibility concerns purely peripheral (i.e., not cognitive) biases, like differences in the morphology of the articulatory organs or sensitivity to rapid acoustic sequences. In this case, somatic morpho-developmental mechanisms would suffice (e.g., Wolpert, 2001). But the more interesting case concerns the biases which involve more complex brain processing, at a cognitive level. These could include a variety of mechanisms, like phonological short-term memory (Section 3.1.4), the processing of signals requiring a fine temporal resolution, fine articulatory motor control, etc. Such mechanisms refer both to the expressive and inductive aspects of language change and can involve both developmental and non-developmental aspects.

The biases can concern:

- the *expressive* process, by biasing fine motor control, speed or activation rules for accessing working memory content, etc., this type of bias can modify the linguistic data which is used for the induction of the next generation's language;
- the *inductive* process, by biasing the probability of attribution of F0 variations to linguistic distinctions, the processing of fast acoustic signals, the activation of items in the working memory, etc., this type of biases can modify the grammar induced from the previous generation's linguistic output,

and can be manifest during:

- *childhood* (the *critical period*), affecting the induction of L_1 by future native

speakers, through a biasing of specific inductive mechanisms (this seem to represent the default mechanism assumed in classical ILM studies, like Kirby, 2001 or Hurford, 2002);

- *adulthood*, affecting either *expression* or *induction* or both. While the expressive biasing is straightforward, the inductive biasing in adulthood can concern either L_2 learning (i.e., adult) -driven language change, including language shift, or the continuous adult L_1 grammar change through usage. These processes can lead to a biased Ostler-type stability through language shifts or to the existence of language change attractors due to adults.

The biological mechanisms involved in adulthood biasing can involve morphological (i.e., selective neural death, axonal growth patterns, etc.) or functional (synaptic behaviour, neurotransmitter or neuromodulator activity, etc.) changes. Therefore, it is important in such discussions not to focus exclusively on L_1 learning children, as the relevant biases might be manifest only in adults.

Concerning the direction of biasing, it can act by *disfavoring* certain values, v_1, ..., v_i, of linguistic feature *f*, or by *favoring* complementary values v_{i+j}, ..., v_n (while values v_{i+1}, ..., v_{i+j-1} are not affected by it). From a purely external point of view, the effects on the trajectory of language change would be the same, affecting the relative probabilities of the two sets of values, but specifically devised experiments can disentangle these competing explanations. For example, in the case of *ASPM*, *MCPH* and *Tone*, it is currently impossible to say that the bias induced by high frequencies of *ASPM* and *MCPH* determine a relative incapacity to acquire tonal distinctions or a relatively increased capacity to use sequential structures[275].

Nevertheless, for the moment, the fact that the exact genetic, molecular, developmental and neuro-cognitive nature of such biases is vague does not affect the arguments supporting the theory of non-spurious correlations. What is important is that they are made increasingly plausible by the accumulating data and theory in the relevant sciences and that they are falsifiable in the sense that specific hypotheses can be formulated in suspected cases and refuted using a scientific approach (Popper, 2002).

275This proves again that simple-minded claims that speakers of non-tonal languages are "superior" can be easily turned on its head (in the same simple-minded manner) and argued that speakers of tonal languages are "superior".

5.1.10. The importance of the theory of non-spurious correlations between genetic and linguistic diversities

Summing up the previous sections, the general theory of non-spurious correlations between genetic and linguistic diversities can be formulated as concerning

> those correlations between inter-population genes frequencies and linguistic variables/features values differences, when the genes are expressed, and there are plausible mechanisms connecting the genes to these linguistic effects.

In more detail, specifically targeted tests must also be performed to systematically study these proposed mechanisms. Moreover, the main process connecting genetic and linguistic differences, in this case, is not represented by demographic effects shaping both diversities in similar ways, but by a process of genetically biased language transmission down the generations, whereby genetic differences, in the appropriate context, bias language expression and/or induction, affecting the trajectory of language change. There can also be a feedback loop from language/culture onto the genes, but this must form the topic of a dedicated investigation.

While theoretically plausible, given what we know from evolutionary biology, genetics, neuro-cognitive sciences and linguistics, this must be backed up by real-world studies and mathematic and computational modeling. If supported by such approaches, it could have a profound impact on the way we understand some language changes and linguistic diversity, as well as, potentially, providing a new tool for studying prehistoric events, combining in a new, and more productive, way linguistic and genetic data. Far from claiming any "new synthesis between linguistics and genetics", and well aware of the dangers posed by them (Section 3.2), these non-spurious correlations have a limited applicability. Nevertheless, their main impact is on the way we understand the interaction between human biology and culture, strengthening the natural link between them.

5.2. Non-spurious correlations and language evolution in the context of human evolution

The fact that during human evolution, the transition between language-less stages and fully-

modern-language-able stages occurred, is hardly disputable, but there is disagreement over most of the relevant details (see, for example, the multitude of opinions expressed during the EvoLang conferences[276]: Hurford, Studdert-Kennedy & Knight, 1998; Knight, Studdert-Kennedy & Hurford, 2000; Cangelosi, Smith & Smith, 2006; or in Christiansen & Kirby, 2003). Simplifying, the main divides seem to concern:

i. the *nature of the transition*: *catastrophic* (e.g., Crow, 2002a, b) versus *gradual/accretionary* (e.g., Pinker & Jackendoff, 2005; Smith, 2006; Corballis, 2004 or Hurford, 2003);
ii. the *timing*: *recent* (e.g., Crow, 2002a, b) versus *ancient* (Hurford, 2003);
iii. the *protolanguage*: *holistic* (e.g., Kirby, 2000 or Wray, 2000, 2002) versus *synthetic* (e.g., Tallermann, 2006, Bickerton 2000),

but few theories address all these aspects in the same detail.

The particular model of human evolution considered strongly constrains the class of language evolution models possible, while there is also a much weaker reciprocal influence from the language evolutionary model onto the human evolutionary model. It must be noted the this remark concerns specifically the relationship between *models* of human and language evolution and not the *process per se*. This is so, in part, due to the difference in data available for model building and testing, whereby human evolution is at a clear advantage compared to language evolution. Given this inter-dependency between these two types of models, it is probably best to talk about *composite human-language evolutionary models*.

As argued throughout Chapter 2, the dominant model of human evolution, by default considered as true, especially outside palaeoanthropology, is represented by author-dependent slight variations on the Recent Out-of-Africa with Replacement[277] motif. This choice of human evolutionary model is usually taken to support, and be supported by a recent, catastrophic (but the actual degree of catastrophism varies) origins of modern language (e.g., see Crow, 2002a, b as an extreme case), forming a composite recent

276The Evolution of Language International Conferences held every two years (1996 Edinburgh, 1998 London, 2000 Paris, 2002 Harvard, 2004 Leipzig and 2006 Roma), http://www.ling.ed.ac.uk/evolang/ (September 2006).

277These slight variations concern mostly the actual dates involved and the (effective) population size (though, rarely made clear that it's the effective population size and not the census population size, see Sections 2.1.1.2 and especially 2.2.3).

catastrophic model (denoted in the following as the *CRC model*). As discussed in Section 3.1.5, the recently identified *FOXP2* gene seems to be the latest killer argument in favor of the CRC, but, for reasons detailed throughout Section 3.1 and especially 3.1.5, this argument is suspect, at best.

But, taking into account the many arguments against ROA discussed in Chapter 2, we are offered the possibility to chose a different type of model, which, even if less pleasant than ROA by being not so categorical and simplifying[278], allows a relaxation of artificial constraints on the language evolution models. This emerging human evolution model, in the vein of Relethford's "Mostly Out of Africa" and Templeton's "Out of Africa again and again" (Sections 2.3.1, 2.3.2), describes a complex and dynamic, meta-population based process, whereby groups interact, both genetically and culturally, over extended geographical, ecological and temporal scales. This, in turn, increases the likelihood of a gradual, accretionary process of language evolution, extended not only in time, but also in space, and fundamentally fueled by genetic and cultural diversity. This composite class of models can be denoted as *CARDD* (*composite accretionary, reticulate and diversity-driven model*) and will be discussed in the following sections.

5.2.1. The *CARDD* class of models

When the artificial requirement for a punctual speciation event is removed from the human evolutionary model (frequently, considered to be the even emergence of modern language – e.g., Klein, 1999; Section 2.2.1), the consequences for language evolution are overwhelming. The concept of a recent speciation event constrains both the temporal span and the amount of diversity (genetic and cultural) available for evolving the modern linguistic capacity, conducing, almost logically, to a brusque emergence, involving either a hopeful monster (*FOXP2*, *protocadherinXY*, etc.) or a more or less purely cultural process[279] (e.g., the emergence of compositionality solely through transmission bottlenecks). But the trouble with both these proposals is that they do not seem to withstand closer scrutiny.

First, no matter how much phenotypic plasticity (especially at the neural level) there might

278 But this is not a scientific argument.

279 Which could account, in principle, for spectacular adaptive changes in a biologically short timespan.

have been in some pre-modern hominids, it is hard to accept that a single lucky mutation would have provided modern language out of something radically different, no matter what the intermediate proposed mechanism is (e.g., Mithen's (1996) "cognitive fluidity", Crow's (2002b) "lateralization", etc.). The difficulty of accepting such accounts does not stem from what Richard Dawkins (Dawkins, 1990a) calls the "Argument from Personal Incredulity", which would reflect the limitations of my own imagination and knowledge, but from a series of solid arguments:

(a) the *biological limits* of phenotypes engendered by mutations seem quite well known[280] and, concerning the non-deleterious effects, which are very rare compared to the deleterious or neutral ones, even when phenotypic plasticity is considered, do not seem able to encompass phenomena of such complexity (e.g., West-Eberhard, 2003; Dawkins, 1990a, 1997; Skelton, 1993; Gerhart & Kirschner, 1997);

(b) the *behavioral genetics of language* (Section 3.1) seems to argue forcefully in favor of an important genetic component, accounted for by many genes with small effects, comprising both generalists and specialists, most of them involved in more than one aspect of language, or, generally, cognition. Moreover, the model of few genes with big effects seems improbable, and catastrophic genes, like *FOXP2*, seem to be aberrant and probably not fundamentally involved in language evolution (Sections 3.1.5 and 3.1.7);

(c) the *indissoluble link between modern humans and modern language*, as argued by various authors on the basis of a specifically modern "package" does not seem to hold. Modernity was certainly not required for *Homo erectus* to expand his geographical and ecological range and reach remote islands (Sections 2.1.2.2 and 2.2.9).

Thus, this type of account seems to have a very low probability.

Concerning the second type of theories, arguing for a purely cultural process (e.g., Kirby, 2000), they still require some form of biological evolution to provide the cognitive processes (potentially, non-language specific) required for a proper cultural evolution of language. But, given the apparently very general requirements (structured meaning space, pattern matching and rule formation, e.g., Kirby, 2000; Brighton, 2003), one is left to wonder if this really requires a modern brain to function. Therefore, the cultural process in language evolution

280But this does not, of course, preclude revolutionary new discoveries.

must also be critically understood in the large context of human evolution as, by themselves, they cannot be used to decide between competing theories.

But, as reviewed above, limiting language evolution in this way raises a series of problems, whose solutions seem to lie outside the single recent species springing from a small population scenario imposed by CRC. In contrast, a CARDD model, by removing the speciation event(s) and allowing temporally, spatially and ecologically large genetic and cultural exchange networks to exist, better matches both the human evolutionary data and also the data relevant for language evolution. Thus, the many genes with small effects suggested by the behavioral genetics of language (Section 3.1), supporting a gradual, accretionary scenario for the evolution of language and, if we consider data from evolutionary biology and genetics (Skelton, 1993; West-Eberhard, 2003), requiring a longer interval than the last 50-150ky, can be easily accommodated by the new timescales. Moreover, it seems that the controversies surrounding the nature of the proto-language can be solved by gradual, incremental changes (e.g., Smith, 2006), much more easily integrated into a CARDD model than into the CRC.

But besides these "accommodative" advantages of CARDD versus CRC, stemming mainly from the different timescales involved (~2my versus 50-150ky) and allowing a better fit for accretionary language evolution models, there are also other, more subtle issues, involving the amount of genetic and cultural diversity required to evolve modern language and the impact of inter-group interactions. Unfortunately, these issues are very rarely (if ever) discussed, on the implicit assumption that diversity is not relevant and can be abstracted away, in order to get to the core, universal processes and properties explaining the emergence of language. Thus, it seems that this non-variationist stance in language evolution is very much akin to the Chomskyan ideal hearer-speaker in an ideal linguistic community (Section 5.1.4).

5.2.2. Genetic and linguistic diversity – the engine of language evolution

A meta-population model, as argued in Chapter 2, involves a dynamic network of populations, expanding, contracting, becoming extinct and being replaced, but continuously

in contact with their neighbors, and, through such local networks, part of regional/global networks of genetic and cultural exchange (Figure 12, Section 2.2.8.2). In such a network of populations, inter-individual and inter-population diversities are essential, and not some sort of noise which must be filtered out in order to gain access to the core, universal properties of interest. Inter-population genetic diversity is fostered by the small size of these populations, increasing the effects of genetic drift (Halliburton, 2004:221-266), as well as the diverse selective pressures due to small-scale (local) and medium-scale (regional) differences in physical[281], ecological[282] and cultural[283] environments. (The cultural diversity is maintained by similar, but less well described, processes.) These levels of diversity allow a more efficient search for adaptations through the processes of biological and cultural evolution, increasing the chance that novel solutions (biological, cultural or a combination thereof) will emerge.

Through the ubiquitous genetic and cultural exchanges connecting these populations, such novel solutions have the possibility to spread on local or even global scales, depending on many factors, including drift (both genetic and cultural) and the properties of the environment(s) in which these solutions prove adaptive. The first one, drift, refers to the random factors conditioning the transmission or not of genetic and cultural innovations, irrelevant to their functional characteristics. The second is more complex and refers to the extent and connectedness of such environments and to the other conditioning factors (environmental, genetic and cultural) affecting the adaptedness of such a novelty. For example, a new genetic or cultural variant might arise, conferring a selective advantage in tropical, humid environments, but its expansion to all such environments of the Old World is conditioned by its ability to cross the arid zones separating them. An interesting case is offered by such variants which prove to be globally adaptive and which could, in principle, spread across the entire range of the species.

Concerning specifically language evolution, the most probable scenario engendered by such a class of models is represented by a very dynamic, two-tired interaction between genetic factors involved in language and its cultural aspects. Let us consider that at time *t*, in a population *P*, there appeared a genetic variant, *g*, biasing the language[284] change in a specific

---

281E.g., climate, Na^+Cl^- availability, UV solar radiation, O_2 levels, etc.

282E.g., disease vectors, predators, food sources, etc.

283E.g., food practices, taboos, etc.

284Language in this context does not refer to modern language but at language as it was at that time, in the respective population.

direction. It might be that *g* is neutral, in which case its fate, and the fate of the language change it induces, is described by genetic drift in meta-population models, eventually leading to extinction or fixation, in the last case, through intermediate stages of inter-populations (and inter-regional) differences.

Or, it might be that *g* is not neutral (on a local, regional or global scale), in which case, it will eventually be eliminated by natural selection (if *g* is deleterious) or fixated in the appropriate environments (in the opposite case). In the last case, if *g* is globally adaptive, it will probably become fixed in the entire species, making also its associated linguistic bias a universal of human language. But a very interesting sub-case is when *g* is non-neutral *because* of its effects on language. For example, it might impair the fine motor coordination of the articulatory organs, generating a deleterious effect, or it might improve the capacity of the phonological working memory, generating an adaptive effect. The actual direction and strength of this selective effect of *g* depends crucially on the linguistic context, but, assuming that it does generate a positive selective pressure, it will tend to spread to neighboring populations. Now, if the linguistic environment in these new populations still determines a positive selective pressure on *g*, it will continue to spread, until, eventually, it will become fixed in the human species, and its linguistic effects, part of the universal linguistic capacity.

While the details of this verbal model need rigorous mathematical and computational modeling, it certainly seems plausible, and would imply a gradual accretion of genetic variants having linguistically adaptive effects, in a certain linguistic context. Of course, by their very spread, these variants change the linguistic context, modifying, thus, the selective opportunities of future genetic variants and insuring a common, universal component of the linguistic capacity. It must be noted that this universality of the linguistic capacity is *not* an assumption, but a *result* of diversity and that it is *dynamic* in time, but not in an absolute progressive manner, towards, say, larger and larger phonological working memories, but in a contextual manner, dependent on history, like any other evolutionary process. Moreover, this temporally dynamic character is manifested by patterning in space, whereby would-be parts of this future universal capacity for language components spread from their origins, across populations, subject to local processes and pressures. This model is very much akin to the theoretical approach of Yamuchi (2004).

Another scenario involves a purely cultural innovation arising at time *t* in population *P*, and which proves adaptive on a local (or even global) scale in the given linguistic context, and which spreads across populations in the same manner that genetic variants do. In this case, the "easiness" with which this cultural innovation arises depends critically on the pre-existing cultural context and, possibly, on the foundation offered by the genetic structure of the population. If this arises "easily", then, it is expected that it will emerge in several centers more or less immediately after the relevant cultural (and genetic) context becomes available, being, in this sense, their immediate consequence. Thus, in such a case, its diffusion is less important compared to its emergence *de novo*. But, if it represents a "hard" to arise innovation, then the diffusional process becomes very important for its future patterning. Nevertheless, it is clear that even for purely cultural innovations, without a direct non-spurious correlation with a genetic novelty, the possible fixations into the universal linguistic repertoire follows a very complex diffusional pattern (possibly from many sources).

Concluding, the inter-population genetic and linguistic diversities, in a meta-population model, determine through a complex dynamics, the emergence of a universal, species-specific biological and cultural[285] linguistic faculty. This is in a continuous process of change, realized through inter-population and inter-regional patterning, and, as any evolutionary process, represents a mosaic of frozen accidents and context-dependent selective pressures. Therefore, it is very probably illusory to search for a unique "core" "essence" of language, like, for example, Hauser, Chomsky & Fitch's (2002) "recursion" (see Parker, 2006a, b, for a thorough critique). This program would be like the patently futile (but still revived with each new generation) search for "the human essence".

5.2.3. A model for language evolution based on inter-population diversity

The model sketched in Sections 5.2.1 and 5.2.2, based on a meta-population approach to human evolution (Section 2.3), critically emerging from inter-population genetic and linguistic diversities connected on a regional/global scale, has focused so far only on the feature/variable side of linguistic diversity. But, as discussed extensively in Sections 3.2.1 and 3.2.2, the other aspect of this process is encapsulated by the descent from a common

285The distinction between biological and cultural is very simplifying and, probably, misleading.

ancestor paradigm, concerning, thus, coherent linguistic groups (dialects, languages, families, etc.), as opposed to their components.

The CARDD class of theories seem to be naturally described, at the linguistic level, by a Dixonian, *punctuated equilibrium*-like model (Section 3.2.2), whereby long periods of linguistic equilibrium, characterized by areal processes, are interrupted by short-term punctuations, characterized by a process of descent from a common ancestor. Nevertheless, this controversial theory must be amended in order to account for the full range of phenomena described by the CARDD.

As alluded to by Dixon himself (Dixon, 1997), the punctuation versus equilibrium distinction is not categorical, but depends on scale. More exactly, as any historical linguist knows, linguistic trees are rarely free of horizontal connections between branches (see; for example, McMahon & McMahon, 2005, for a very clear description of the problem and the proposed solutions). For example, the Balkans are a long-time recognized *Sprachbund* (e.g., Thomason, 2000; Joseph, 1999; Tomić, 2003), where areal effects cross the boundaries between four branches of the Indo-European linguistic family: Italic (Romanian), Greek (Greek), Albanian (Albanian) and Slavic (Bulgarian, Macedonian, etc.). The shared features comprise phonetics (e.g., [ə]), lexical items (e.g., "donkey", "box"), morphology (e.g., postposed definite article) and syntax (e.g., general replacement of infinitives by subjunctives). Thus, the two processes, of equilibrium (illustrated by the sharing of features) and punctuation (e.g., the differentiation of south Slavic) coexist on different temporal and geographical scales.

This is an expected characteristic of any CARDD situation, where, on a small scale (both temporal and geographical), neighboring populations can be in an equilibrium situation, ruptured by linguistic replacements (due to the original population shifting to a new language or by demographic replacement because of extermination or "natural" extinction[286]). These small-scale processes can be part of regional-scale equilibrium states or small-scale punctuations could escalate to a superior scale, when, for example, a particularly successful group manages to conquer a larger area or, when large-scale fluctuations (climate, disease, earthquakes, etc.) give one group the opportunity to expand at the expense of the

286Climatic catastrophe, disease, sex-ration fluctuations in a small population, etc.

neighboring ones. It must be noted that the probability of large-scale punctuations probably decreases non-linearly with the size of the concerned area, so that continental or global punctuations become very unlikely. Nevertheless, if one considers the entire span of human evolution, encompassing ~2my, then it seems probable that such large-scale punctuations did happen. Another note concerns the fact that the linguistic entities concerned by such large-scale punctuations spread with modification, due to language shift or natural language change. The process described must be modeled in detail so that its exact characteristics and generated patterns can be studied, but an interesting feature seems to be its self-similarity at different scales, both spatial and temporal. Also, the high dynamism of climate during the last 2my must be taken into account when formulating the exact nature of the equilibrium/punctuation events, avoiding the impression of long-term stability (see above and Section 3.2.2).

As argued previously (Chapters 3 and 4), it is to be expected that different linguistic features have different temporal "stabilities", in the sense that they tend to survive more or less unchanged through successive language shifts. Coupled with non-spurious correlations, this would confirm, in principle, Johanna Nichol's program (Nichols, 1992) of using typological patterns as indices of ancient demographic processes. For example, it is conceivable that the current distribution of linguistic features reflects (across language shifts) ancient patterns of successive waves across the Old World, originating in different places and affecting different features (some through non-spurious correlations with genetic variants), some fixated and some not.

In conclusion, a CARDD model seems to suggest a Dixon-like model for language evolution, amended to account for different stabilities of the linguistic features and non-spurious correlations with genetic variants, which also allows tree-like patterns, based on descent from a common ancestor, all showing self-similarity at different temporal and spatial scales. Moreover, the tree pattern is superimposed by a network of linguistic feature similarities due to different stabilities through language shift, borrowability and non-spurious correlations. Thus, for example, a very stable linguistic feature will survive language replacement, while a very easy to borrow one will cross the boundaries of this new replacement. Therefore, the tree and wave models (e.g., McMahon & McMahon, 2005) must be supplemented with a trans-language shift survival model, eventually due to non-spurious

correlations.

5.2.4. The case of Scandinavian languages: a refinement of the theory

As a concrete example for the previous discussions, let us assume as proven that *ASPM*, *MCPH* and *Tone* are non-accidentally correlated in the sense discussed in Section 4.9. Then, before the emergence of the derived haplogroups of *ASPM* and *MCPH* (before 60kya), all languages were using tonal distinctions[287]. Assuming that the positive selective pressures for these derived haplogroups of *ASPM* and *MCPH* were approximately constant during the last 100ky or so[288], then, after the appearance of *MCPH-D* (~37kya) and with its subsequent increase in frequency, the probability of linguistically using tone distinctions decreased towards ~0.5, and continues decreasing[289] after the appearance of *ASPM-D* (~5.8kya) towards 0. Thus, in this sense, the areal presence of tone distinctions, cutting across linguistic family boundaries, is to be attributed not (only or principally) to borrowing, but also to the sharing of the genetic bias towards tonality, due to specific population frequencies of *ASPM-D* and *MCPH-D*. But, it must be noted that the progression through time sketched here is not absolute and irreversible, but intrinsically *statistical* in nature. A very good illustration is represented by the case of *pitch accent systems* in some near-Baltic languages.

Hirst & Di Cristo (1998) offer a good overview of the various intonation systems found in different languages and, in this context, approach the case of the somehow intermediate cases between stress and tone systems. As they say:

> It has been suggested [...] that the classical typological distinction between **stress** languages and **tone** languages should be extended to a three-way distinction between **stress** languages like English, Dutch, Russian *etc.*, sometimes called "dynamic stress" languages or "stress-accent" languages, **tone** languages (like Chinese, Vietnamese and Thai) and **pitch accent** or **tonal accent** languages (like Japanese and perhaps Swedish) (Hirst & Di Cristo, 1998:9, **bold** and *italic* in original),

but it is too early to rule out a continuous classificatory scale (Hirst & Di Cristo, 1998:9).

287This is a very simplifying assumption, neglecting all the other possible factors. Nevertheless, it seems also suggested by independent considerations connected to the usage of F0 for emotional signaling.

288Which, of course, might be false. It might be that the positive selective pressures are very recent, postdating agriculture.

289This, again, must not be taken to mean that tonal languages are somehow "primitive"!

Concerning specifically Japanese and Swedish, they consider that "[...] both appear to possess characteristics of both paradigmatic (tonal) and syntagmatic (accentual) prosodic systems" (Hirst & Di Cristo, 1998:11), and Table 42 reproduces their typological classification based on lexical prosody (Hirst & Di Cristo, 1998:12, Table 1).

Type	*Example*	*Number of lexical tones*	*Lexical stress*
Fixed stress	*Finnish*	0	No
Free stress	*Greek*	0	Yes
Accentual tone	*Japanese*	1	No
Tone Accent	*Swedish*	1	Yes
Tone	*Thai*	>1	No
Tone and stress	*Chinese*	>1	Yes

Table 42: Typological classification based on lexical prosody.
Reproduced from Hirst & Di Cristo (1998:12, Table 1).

Thus, as previously hinted (Sections 4.2.3 and 5.1.2), there are some intermediate cases between the canonical tone and non-tone languages, illustrated by Japanese and, more importantly in the current context, by the Scandinavian languages.

The case of the well-known Scandinavian pitch accent languages/dialects proves to be, in fact, extremely complex. As discussed, for example, by Koptjevskaja-Tamm (2006), the phenomenon of *polytonicity*, defined as "[...] the existence of tonal suprasegmental oppositions in a language" (Koptjevskaja-Tamm, 2006:9) represented

> [...] the original impetus for talking about a Sprachbund in the circum-Baltic areas [including] Norwegian (except for an area in the west), most Danish dialects, Swedish (apart from most of the dialects in Finland and in the neighbouring areas and in Estonia), some Low German dialects, Northern Cashubian [Ethnologue [csb]], Lithuanian, Latvian, Livonian and Estonian (Koptjevskaja-Tamm, 2006:9),

where

> [t]onal phenomena in the CB [i.e., Circum-Baltic] languages are of the "word accent", "lexical accent" or "pitch accent" type, as opposed to word tones [in canonical tone languages] [...] here, the choice between accents is made only once in each word, whereas in tone languages, (almost) every syllable has its own tone (Koptjevskaja-Tamm, 2006:9).

So, this phenomenon is not restricted only to Scandinavian languages, but concerns some more cases. Nevertheless, the pitch accent in the Circum-Baltic area does not represent a

unitary phenomenon, and can be divided in three distinct groups[290] (Koptjevskaja-Tamm, 2006:9). The first group is represented by the Baltic languages, in which this is a residue of a

> [...] much wider phenomenon once covering large parts of the Indo-European dialect area [and] genetically related to polytonicity in Slavic languages (which still exists in certain varieties of Slovene and Serbo-Croat) and in classical Greek (Koptjevskaja-Tamm, 2006:10).

This original system, reflected in the Baltic languages, is summarized by Fortson (2004):

> From the available comparative evidence, it is standardly agreed that PIE [i.e., Proto-Indo-European] was a pitch accent language. There are numerous indications that the accented syllable was higher in pitch than the surrounding syllables (Fortson, 2004:62).

Its current reflexes are (Koptjevskaja-Tamm, 2006:10-11), for example, in Lithuanian (2 tones, acoustically variable across dialects) and Latvian (2 or 3 tones; originally 3 but reduced to 2 in most dialects, the 3rd tone involves a glottal closure "broken tone" or "Stoβton").

The second (and best-known) group is represented by the Scandinavian languages, where pitch accent is "supposed to be a relatively recent phenomenon found in most dialects of Norwegian, Swedish (except in dialects in contact with Finnish, Saami and Estonian), including Dalecarlian [Ethnologue [dlc]], and in Danish – but not in Icelandic or Faroese" (Koptjevskaja-Tamm, 2006:11), but their origins and historical developments are controversial (see below). Even in this area, polytonicity is not uniform as "[w]ord accents par excellence (i.e., tonal accents) are found across Norwegian and Swedish, while Danish has an opposition between syllables with and without a glottal closure, **stød** [similar to "Stoβton"]" (Koptjevskaja-Tamm, 2006:11, **bold** in original), and they cannot be related to the PIE pitch accent system, representing, thus, *de novo* innovations (Koptjevskaja-Tamm, 2006:11).

The third group comprises the Finnic languages Estonian and Livonian, where polytonicity represents "[...], again, a different phenomenon, which has to do with the reduction of non-initial syllables (and, ultimately with the fixed initial stress) and a compensatory secondary lengthening of the initial syllable, or overlength" (Koptjevskaja-Tamm, 2006:12).

There are also some other languages/dialects which show polytonicity but are not in the Circum-Baltic area, forming a fourth group, composed of several West-Germanic dialects

290Concerning specifically this Circum-Baltic region.

spoken in the Cologne-Trier area (West Germany, Eastern Netherlands, Belgium, Luxembourg), showing a distinction between *push-tone* ("Stoßton, Schärfung"/"valtoon, stoottoon") and *drag-tone* ("Schleifton"/"sleeptoon"). This phenomenon is known as the *Rhineland Accentuation* ("Rheinische Akzentuierung") (Koptjevskaja-Tamm, 2006:13). The German dialects form two groups, *Rule A* and *Rule B*, which show opposite distributions of tones, while the Dutch dialects seem to have developed similarly to German *Rule A* (Koptjevskaja-Tamm, 2006:13). These dialects "[...] differ crucially from Scandinavian in having tonal oppositions in monosyllables" (Koptjevskaja-Tamm, 2006:13), and there does not seem to exist any direct relationship between them.

Overall, it seems that these three (four) groups of languages showing polytonicity are independent, in the sense that "[...] there are no obvious connections among [them]" (Koptjevskaja-Tamm, 2006:14), and, their histories are more or less independent, even if, through a process potentially very interesting for the argument of this thesis, "[...] an incipient internally motivated linguistic change in a language may be reinforced by contacts with another language that either shows the "target" characteristics of such a change or is moving in the same direction" (Koptjevskaja-Tamm, 2006:14).

Concerning specifically the Scandinavian languages, the situation is very complex, both synchronically (e.g., Koptjevskaja-Tamm, 2006; Kristoffersen, 2006, 2003; Bruce, 2004; Riad, 1996 or Bye, 2004) and diachronically. Functionally, pitch accent is not very important in the Scandinavian languages (Bye, 2004:3), with some 2400 minimal pairs for Norwegian and some 350 for Swedish (Bye, 2004:3-4). The related problems of the origin and spread of pitch accent systems in the Scandinavian languages are very contentious and far from solved (e.g., Koptjevskaja-Tamm, 2006; Bye, 2004; Riad, 1998). Nevertheless, it seems agreed that they represent relatively recent phenomena, without a direct connection to the PIE pitch accent system (Koptjevskaja-Tamm, 2006:10). A recent origins theory, positing an *Old Scandinavian* (1000-1200 AD or earlier) origins for pitch accent, offers the following explanation: "[...] words which were monosyllabic in Old Scandinavian have reflexes with Accent 1, whereas words which were polysyllabic in Old Scandinavian have reflexes with Accent 2" (Bye, 2004:10), while an older origins theory, places the relevant events during the *Proto-Nordic* period (800-850 AD) (e.g., Bye, 2004; Koptjevskaja-Tamm, 2006; Riad, 1998). It appeared as an attempt to account for some phenomena unexplainable by the Old

Scandinavian theory, and attempts to trace them back to the Syncope Period during Proto-Nordic, where "[...] words which lost a medial syllable in the Syncope Period acquired (or retained) Accent 1, while those which did not lose a medial syllable acquired (or retained) Accent 2" (Bye, 2004:10), through a process of *stress clash resolution* (Bye, 2004:11, 42-47; Koptjevskaja-Tamm, 2006:12; Riad, 1998). Nevertheless, "[i]t is widely agreed that the phonemic opposition itself arose [recently], but that the pitch differences underlaying it had been around for a considerable time" (Koptjevskaja-Tamm, 2006:11), but, with the important caveat that "[...] recognising non-distinctive tones ("singing intonation" in the distant past of languages is hardly that simple" (Koptjevskaja-Tamm, 2006:19).

While this second hypothesis seems the most currently accepted (e.g., Bye, 2004; Koptjevskaja-Tamm, 2006; Riad, 1998), there are still controversies, and counter-proposals and amendments are being made. For example, Bye (2004) argues for a different process (pitch target delay), which would reverse the ancient versus recent classification of the concerned dialects, while Koptjevskaja-Tamm (2006) criticizes the classic notion that the processes accountable for the origins of pitch accent were "[...] applied to a basically uniform language - "Proto-Nordic" - and yield another, which however either immediately (partly via those changes themselves) or shortly afterwards split up in two dialects: "West Nordic" (Norwegian, Icelandic) and "East Nordic" (Danish, Swedish)" (Koptjevskaja-Tamm, 2006:15), and argues that the apparent linguistic uniformity of Scandinavia resulted from a recent, gradual language shift, due to the spread of a/some Danish dialect(s) (Koptjevskaja-Tamm, 2006:15-17). She argues that "[...] the lexical distinctions expressed by tone accents originated somewhere in Denmark and spread from there by means of the prestige language" (Koptjevskaja-Tamm, 2006:17), while this original tonal distinction in Danish was later replaced by stød (Koptjevskaja-Tamm, 2006:12).

Nevertheless, the main problem faced by the early origins account (Proto-Nordic) is that neither Faroese nor Icelandic show such phenomena, and, given that they were colonized during the 9th - 10th centuries, would favor the most parsimonious hypothesis that, by that time, there was yet no tonal distinction available in the Scandinavian languages (or, at least, in West Norwegian dialects). But Koptjevskaja-Tamm (2006:18) argues that the correct explanation is given by the subsequent loss of pitch accent in Icelandic and Faroese through contact processes, similar to the Swedish dialects in contact with Finnish and Estonian.

To conclude, from our perspective, the following points are relevant:

i. there are several, apparently not directly related (but possibly mutually influencing) pitch accent varieties around the Baltic sea and in North-Western Europe;

ii. in the Baltic languages case, these represent a continuation of early Proto-Indo-European phenomena;

iii. in the Scandinavian languages, the origin is fairly recent (the last 2ky) and internal;

iv. the West Germanic dialects phenomena are not related to the Scandinavian cases.

While the frequency of the derived haplogroups of *ASPM* and *MCPH* is unknown for these populations[291], it seems safe[292] to assume that they do not deviate too much from the European pattern of high *ASPM-D* and *MCPH-D*. In this case, our theory would predict a strong bias against tone distinctions/towards sequentiality, signaling an apparent paradox and a possible falsification of the theory.

The first observation is that, in a strictly binary classification into tone and non-tone languages, most pitch accent systems should probably go with non-tone languages, given the relative functional non-importance of pitch accent (e.g., the Scandinavian languages) and the massive difference in tone distinctions compared to canonical tone languages. Nevertheless, such arguments are always subjective and open to criticism. The second, and most relevant argument, is that even if one would allow a more refined scale (e.g., Maddieson, 2005), the ordering would be

tone > pitch accent > non-tone

Nevertheless, with this observation, the statistical nature of the non-spurious correlations theory, as applied to tone, becomes evident: high frequencies of *ASPM-D* and *MCPH-D* in a population bias the appropriate language(s) away from using tone distinctions, but not in an absolute way. Therefore, limited uses of tone distinctions (i.e., pitch accent) can appear or can be maintained in the right circumstances, of which contact and mutual reinforcement with other languages in the same stage (pitch accent) probably represents the most important one.

More precisely, the generic Indo-European trend seems to be to lose the old PIE pitch accent system, except in some cases, and this trend can be explained naturally by the theory of non-

291And a very interesting direction of future research.
292But there could be some surprises concerning Finns, Saami and possibly even Scandinavians.

spurious correlations. The age of PIE is approximated usually at some 4-6ky (Mallory, 1991; Fortson, 2004)[293], time around which, presumably, *ASPM-D* appeared and spread, lowering the probability of using tone distinctions (also referring to pitch accent). Thus, the IE general trend of losing pitch accent can be related to the increase in frequency of *ASPM-D* (in an already assumed context of high *MCPH-D*). In the context of this general IE trend, the case of the Baltic languages, maintaining the reflexes of the old PIE pitch accent system, can be understood. However, the tonal system of these languages does not seem to increase in complexity, the opposite being apparently true (i.e, the collapse of the original three tones system of Latvian into two tones, Koptjevskaja-Tamm, 2006:10). Moreover, the Scandinavian and the unrelated West Germanic cases point to a recent innovation, whereby sound changes gave rise to a limited usage of tone distinctions.

If the above interpretation of the Icelandic and Faroese cases is true, then we also have the fact that these systems are unstable when in situations of contact with languages not using tone distinctions (also Swedish in contact with Finnish and Estonian). Moreover, considering Koptjevskaja-Tamm's (2006:14) mutual reinforcement observation, there emerges a picture whereby high frequencies of *ASPM-D* and *MCPH-D* decrease the probability of using tone distinctions, which is manifested in two (diachronically related) ways:

(a) a decrease in the probability of tonal languages;

(b) a decrease in the maximum complexity of tone distinctions attainable by such languages, when tone distinctions do happen (or persist from previous stages).

Therefore, it is postulated that (b) represents a valid mechanism for (a), and that (b) is a direct consequence of the theory of non-spurious correlations. Thus, for illustration purposes only, let us consider a population P, which, at time t_0, has low frequencies of both *ASPM-D* and *MCPH-D* ($v_{ASPM,0}$, $v_{MCPH,0} \approx 0$), allowing a canonical tone language, L_0. At a subsequent time $t_1 > t_0$, the frequencies of both haplogroups increase into the non-tone regime ($v_{ASPM,1}$, $v_{MCPH,1} \approx 1$) and the language gradually changes into a pitch accent one, L_1. As time goes by, $t_2 > t_1$, the frequencies remain constant ($v_{ASPM,2}$, $v_{MCPH,2} \approx v_{ASPM,1}$, $v_{MCPH,1}$) and the language, L_2, can either remain in a low-complexity pitch accent state, especially if reinforced by contact with other pitch accent languages, or continue to change until reaching a canonical non-tone state. This idealized scenario is represented in Figure 73, and is reminiscent of the Baltic

293But see also the very controversial estimates of ~10kya (Renfrew, 1991; see also Mallory, 1991, Fortson, 2004 for comments).

languages.

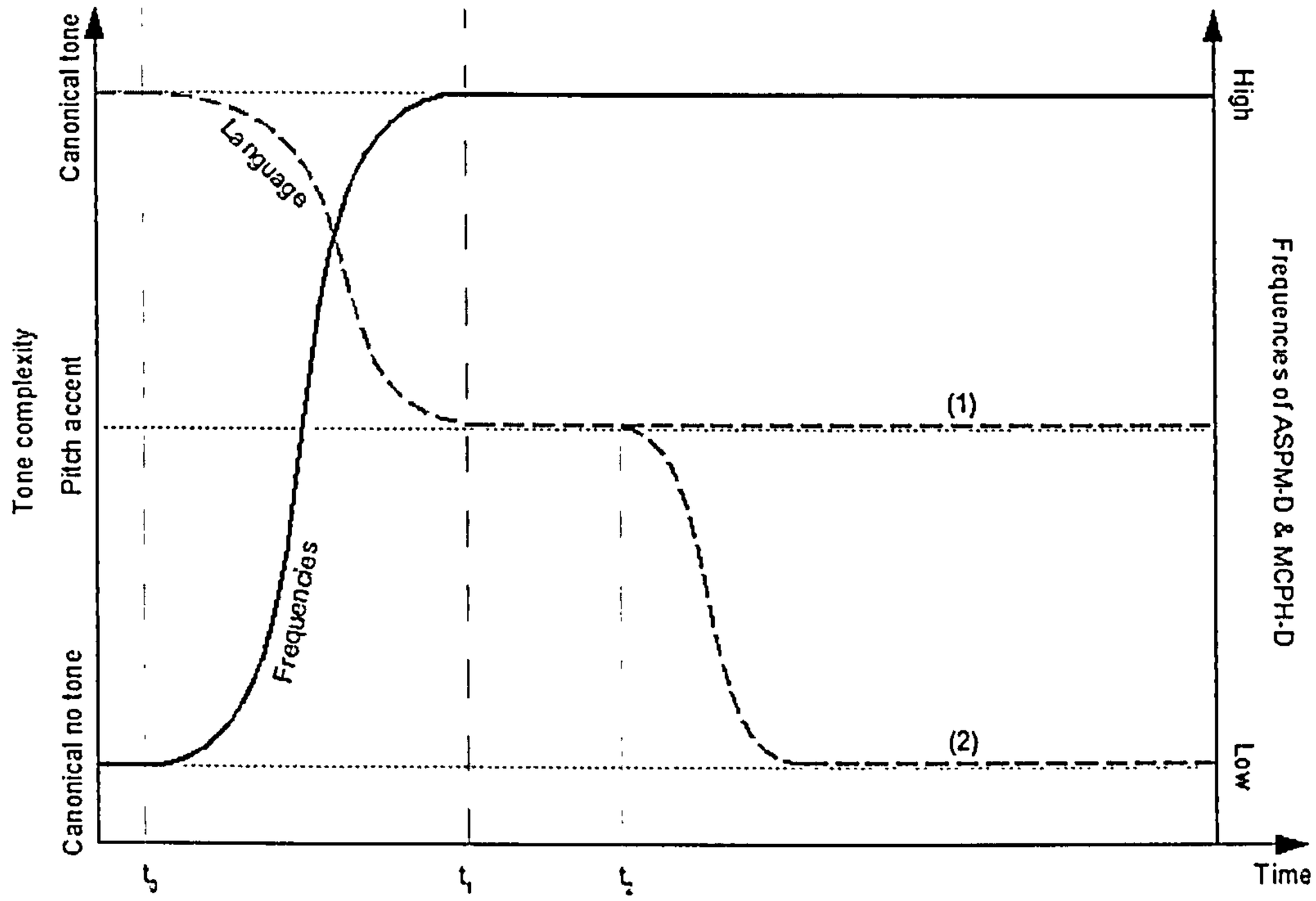

Figure 73: The idealized behavior of the language when the frequencies of *ASPM-D* and *MCPH-D* increase.

Solid black line = frequencies of ASPM-D and MCPH-D (considered synchronized), gray interrupted line = language trajectory. At time t_0, ASPM-D and MCPH-D have low frequencies, allowing a canonical tone language, then the frequencies of ASPM-D and MCPH-D increase until they approach 1 at time t_1. During this process, the language becomes less and less tonal, tending towards a pitch accent language. Then, there are two possible trajectories: (1) the languages conserves pitch accent (presumably in contact with other such languages, through mutual reinforcement) or (2) the tonal distinctions collapse and the language becomes a canonical non tone language.

Another scenario appears for a population *P*, constantly with high frequencies of *ASPM-D* and *MCPH-D* (previous scenario after t_2, trajectory (2)). Let us suppose that at time $t_3 > t_2$, due to language-internal processes (or other factors, like contact or language shift), the population *P* initiates a tone distinction, which evolves towards pitch accent. But the theory predicts that the probability that the language will continue to increase the complexity of its tone system is very low[294], while the most probable outcomes are either the reversal towards no tones or the maintenance of the pitch accent in the right circumstances (like contact with other such languages, through mutual reinforcement). This idealized process, reminiscent (and predictive) of the Scandinavian case, is represented in Figure 74.

294As opposed to the case where *ASPM-D* and *MCPH-D* have low frequencies.

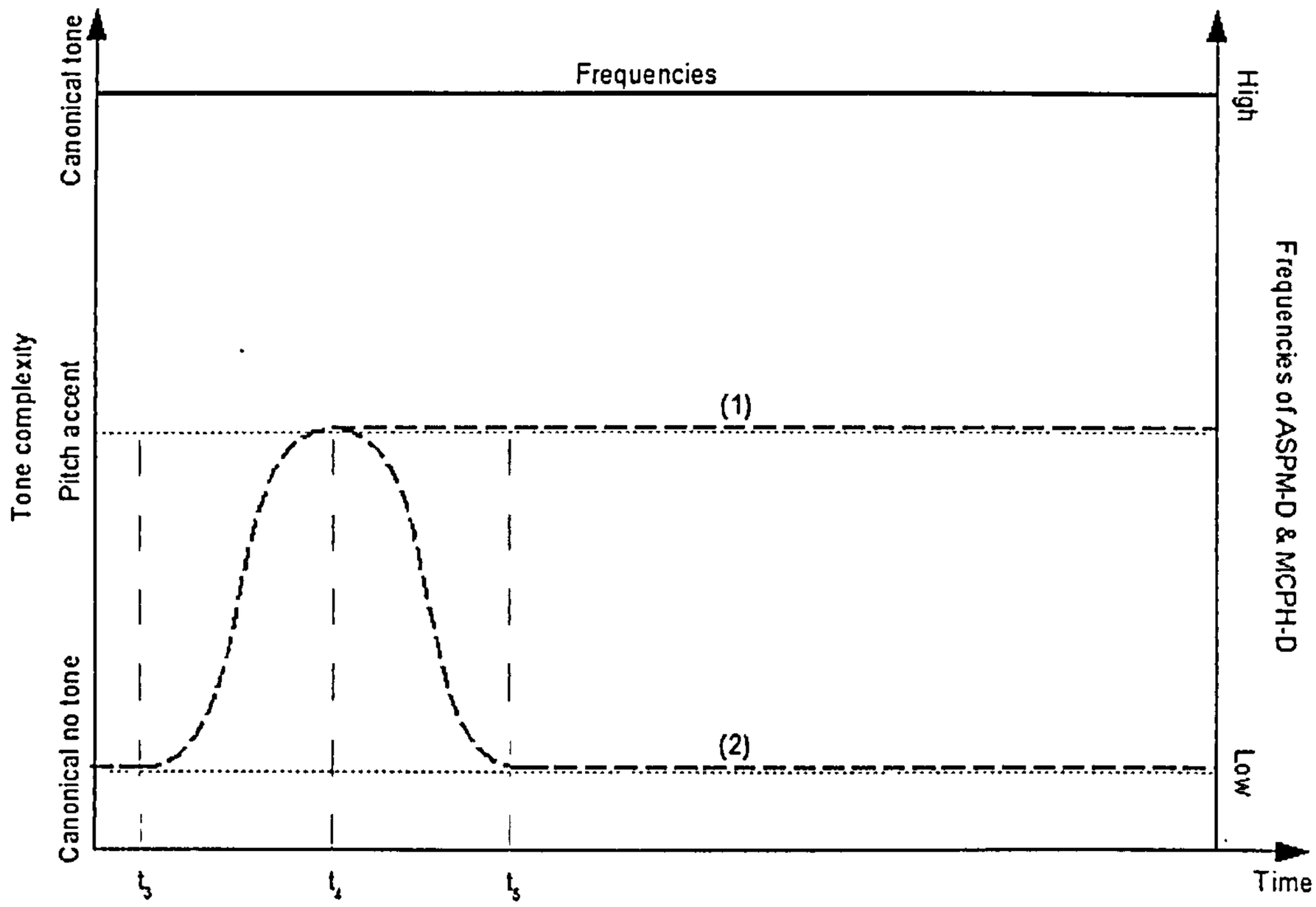

Figure 74: The idealized behavior of the language when the frequencies of *ASPM-D* and *MCPH-D* are constantly high.

Solid black line = frequencies of ASPM-D and MCPH-D (considered synchronized), gray interrupted line = language trajectory. At time t_3, a fluctuation (internal language change, etc.) initiates a process of tone distinction, which increases towards a maximum of complexity (pitch accent), after which is can persist in the right circumstances (1) or revert to a canonical no tone language (2).

It must be highlighted that these scenarios are highly idealized and that, in reality, they are intrinsically probabilistic[295] and dependent on context (internal language factors, language contact, etc.). The rigorous estimation of these probabilities will require detailed mathematical and computational models, as well as calibration with real data. Nevertheless, this model allows a further refinement of the non-spurious correlations theory in the case of *ASPM*, *MCPH* and *Tone*, which makes the following new predictions:

- it is possible that the pitch accent system of PIE represents the collapse of a previous canonical tonal system, following the increase in frequency of *ASPM-D* (after its appearance, ~ 5.8ky, 95% CI: 0.5-14.1ky) in the context of high frequency of *MCPH-D*;
- the current pitch accent systems of Baltic, Scandinavian and West Germanic dialects will most probably not evolve towards canonical tone systems;

295I.e., the trajectories will oscillate around their local stable values.

- in any region with high frequencies of *ASPM-D* and *MCPH-D*, given enough time for the bias to influence language change, the most frequent languages will be canonical no tones, with some pitch accent cases, mostly forming reciprocally reinforcing linguistic areas.

5.2.5. The distribution of ages of the non-spurious correlations

The non-spurious correlations between genetic and linguistic diversities involve, on one hand, genes which have (however indirect) biasing effects on language, and the relative probabilities of linguistic feature values, on the other hand. As previously argued (Section 4.9), the temporal stability of linguistic features is not uniform, with some being very unstable and some better capable of resisting language change, contact or shift. The same applies also to the genetic side of the relationship, with some genes being recent and others more ancient[296].

In the particular case covered by this thesis, both *ASPM-D* and *MCPH-D* are relatively recent compared to the span of human evolution, but still old (*ASPM-D*, ~5.8kya) and extremely old (*MCPH-D*, ~37kya) by linguistic standards. Thus, if the theory of non-spurious correlations in general, and this particular case, in special, will resist further tests, then *ASPM-D*, and, especially *MCPH-D*, carry information about ancient linguistic matters, far more ancient than the comparative method, or any of the proposed "non-orthodox" approaches[297], can dream to reach.

But it is certainly possible, in theory, to think that the actual age of the gene is irrelevant in the context of such a non-spurious correlation. It could be possible, for example, to discover a non-spurious correlation between an allele, *A*, originating 150kya, 500kya or even 1.5mya and a certain linguistic bias. But how must we interpret such a relationship? It could be the case that the current relationship is also the original one, the same biasing effect persisting unchanged, in which case, we would have access to a source of information concerning extremely ancient linguistic states. But this will depend on the exact nature of the biasing effect and how "modern" this is judged to be. Another equally plausible case is that the

296It must be noted that, when speaking about the "age of a gene", this actually represents a shorthand for the age of appearance of the gene's relevant allele, not the locus itself.

297Mass comparison, macro-families, etc. (see Section 3.2).

biasing effect changed with the changing genetic and linguistic context, meaning that the allele biased language in a certain way originally, then in a different way in a new context, and, finally, in another different way in the current context. In this case the current non-spurious correlation cannot be used to infer much about very old linguistic states. A third possibility is that the linguistic biasing effects of allele *A* were not originally present, but were acquired (much) later, when the genetic and/or linguistic contexts changed appropriately. In this case, the non-spurious correlation still carries information concerning old linguistic states, but not contemporaneous with the allele's appearance. Given that, in particular cases, a combination of these scenarios is possible and that more than a single gene (locus) are involved, potentially affecting more than a single linguistic feature, any interpretation must be carefully weighted by the actual evidence. Most importantly, the age of the relevant allele(s) represents only a lower limit for the age of the linguistic phenomenon, which might be much more recent.

The human evolutionary model profoundly impacts the age profile for such non-spurious correlations, in the sense that ROA, through its postulated speciation bottleneck, would effectively limit the age of such correlations to the age of the bottleneck[298], while a meta-population model of the type advocated in Chapter 2 would allow a much larger pool of variation, with alleles of different ages and different regional origins accreting into the living population (at any moment in time). Therefore, it would allow the existence of very ancient non-spurious correlations, but this must remain a purely theoretical conjecture waiting for future research.

5.3. Conclusions and future directions

The main argument of this thesis is that genetic and cultural diversities, far from representing a nuisance and a potential for political mistreatment, are one of the most important features of our species, allowing us to evolve in the first place, and to spread across the Earth by adapting to almost every available niche, and creating new ones.

Concerning our origins (Chapter 2), the apparently best supported human evolutionary model, the Recent Out-of-Africa with Replacement, turns out to fail to account for all the

298This is still a probabilistic effect, but most of the ages must cluster around this date.

currently available data, while findings usually taken to support it against its competitors, are, on closer scrutiny, unable to actually distinguish between them. Moreover, it seems that, at least partially, its impetus is not purely scientific in nature, but also political, as it is sometimes taken to champion a distorted and uniformity-driven account of human nature, covering under the blanket the differences in the name of the current political correctness. Surprisingly, it appears that such a radical monogenism, founded on speciations, splits and divergence is more prone to racist musings than the alternative, diversity-based approaches. These alternatives are based on a concept of the human species as composed of inter-connected populations, fostering diversity and, at the same time, evolving together, intrinsically the same.

That there is a relationship between genetic and linguistic diversities is obvious, but this relationship is not unitary, being composed of various aspects (Chapter 3). Inter-individual differences in the genetic makeup account for an important proportion of the inter-individual differences in various linguistic aspects and shed light on the complex relationship between genes and environment in shaping the linguistic phenomena. They also help disentangle the various aspects of linguistic abilities and disabilities, their co-occurrence, the nature of these genetic influences and the most probable model accounting for them. It is concluded that a many genes with small effects model, as opposed to a few genes with large effects model, accounts better for the available data and that catastrophic effects, such as those associated with *FOXP2*, while very interesting, do not shed much light on these general mechanisms. Another aspect concerns the relationship between inter-population genetic diversity and linguistic diversity, both viewed as deriving from common historic demographic processes. This type of correlations was very acclaimed in the recent past as providing a "new synthesis" between genetics, linguistics and archaeology, as offering new keys to understanding history and pre-history. Unfortunately, it turns out that some of its fundamental assumptions do not hold, and the promised synthesis is more of a uni-directional program, whereby geneticists (and archaeologists) elect and force linguistic theories into their pre-existing models, without paying much attention to the linguists themselves. Therefore, while there are very interesting results and promising techniques, the field is very exposed to abuses, and it has generated very strong negative reactions in the three communities involved, making a real joint program much more difficult.

But there exists another way of looking at linguistic diversity, based on the idea of linguistic variable or feature, and which regards linguistic diversity not only from a common-ancestor perspective, but also considering contact and genetic influences. Therefore, the hypothesis that there exists a special correlation between two genetic variants involved in brain growth and development and linguistic tone is tested in a Old World sample of 49 populations, using data for 983 genetic variants and 26 linguistic features, and controlling for geography and history (Chapter 4). This hypothesis cannot be rejected, allowing the theorizing of the existence of non-spurious correlations between genetic and linguistic diversities, whereby genetic differences produce slight biases, which, through an iterated learning-like process, modify the trajectory of language change.

The assumptions, meaning and impact of this novel theory was analyzed in detail in Chapter 5, where the links between the model of human evolution, non-spurious correlations and language evolution are also explored. The main conclusion is that genetic and linguistic diversities, connected through such non-spurious correlations, represent probably the only plausible mechanism for explaining a gradual evolution of language, as opposed to catastrophic mutations or poorly specified purely cultural models. Therefore, this diversity represents not just some sort of noise but the even engine of human and language evolution itself, evolution with continues today, irrespective of our theories, desires and, most often, mis-targeted attempts at harnessing it in the most fashionable directions of the day[299].

The theory of non-spurious correlations, which involves a certain fracture with the previous paradigm of thinking about genetic and linguistic diversities and their inter-relationships, even if supported by the data and techniques developed in this thesis, must be further tested in various ways (including statistical, experimental and mathematical and computational modeling). If one of these attempts will successfully reject it, then science will note that this direction is wrong, and it will be extremely useful to know why such a theoretically viable approach does not apply in practice. But if it will not be falsified and will resist the test of time, then its consequences for linguistics, (pre)history and human evolution will probably be fundamental. Also affected will be the way we conceptualize human diversity and,

299Eugenics is the predilect example of such an attempt (e.g., Wolpoff & Caspari, 1997), but who knows how our grandchildren will look upon our current attempts at making everybody the same, in the image of our own society and values? I think they will probably despise us profoundly for the unequaled destruction of diversity currently taking place, and morally justified in innumerable ways.

hopefully, we will be able to finally go beyond the current policies designed only to *cope* with it.

Annex 1: An overview of the Most Recent Common Ancestor (MRCA), coalescence theory, gene genealogy and expected coalescence time

Any non-recombining DNA lineages, like mtDNA or the Y chromosome (NRY)[300], in any individual in any generation will have exactly one source. Let us focus on mtDNA and consider a constant-size population evolving through time, depicted in Figure 75: the vertical axis represents time in generations, flowing downwards. The horizontal axis is non-directional and represents simply a column identifier (position). There are 12 generations, each with 9 individuals, each individual being either a female (circle) or a male (triangle) and uniquely identified by a pair (generation, position). For example, the bottommost, leftmost individual (a male) is (9,0), while the topmost dark gray female is (0,4). Simplifying, an individual (irrespective of sex) will be denoted $i_{g,p}$, where *g* is its generation and *p* its position; females are $f_{g,p}$, and males, $m_{g,p}$. The arrows represent parental relationships, flowing from one generation to the next: the gray arrows represent paternal relations and the black arrows, maternal relations (mtDNA flows from mothers to their children). Individuals colored white fail to reproduce, while light gray individuals fail to contribute mtDNA into the last generation (11).

Let us consider all the individuals in the last generation: $m_{11,0}$, $f_{11,1}$, $f_{11,2}$, $m_{11,3}$, $f_{11,4}$, $m_{11,5}$, $m_{11,6}$, $f_{11,7}$ and $m_{11,8}$; they are colored in black ($m_{11,0}$, $f_{11,1}$, $f_{11,2}$, $m_{11,3}$, $f_{11,4}$) and dark gray ($m_{11,5}$, $m_{11,6}$, $f_{11,7}$ and $m_{11,8}$), respectively. This difference in color could be interpreted as either a sampling procedure (a study which samples only the black-colored individuals and nothing else) or a real population structure (the inhabitants of different continents).

Each individual living in the last generation (11) inherits its mtDNA directly from its mother in generation 10. Thus, $m_{11,0}$ and $f_{11,1}$ inherit their mtDNA from $f_{10,0}$; $f_{10,2}$, $m_{10,3}$, $f_{10,4}$ from $f_{10,3}$ and $m_{10,5}$, $m_{10,6}$, $f_{10,7}$, $m_{10,8}$ from $f_{10,7}$. It can be seen that, already in one generation, the number of mtDNA lineages has been reduced from 9 to 3 ($f_{10,0}$, $f_{10,3}$ and $f_{10,7}$). The other individuals in generation 10 have failed to transmit their mtDNA into the current population, either because they failed to reproduce at all ($f_{10,2}$, $m_{10,5}$ and $m_{10,8}$) or because they were males ($m_{10,1}$, $m_{10,4}$ and $m_{10,6}$) – note that these males did reproduce and their nuclear genes are represented in the

300Containing a non-recombining portion (*NRY*, more than 90% of its length) and a short *pseudoautosomal* region, which recombines with the homologous regions of the X chromosome.

current generation.

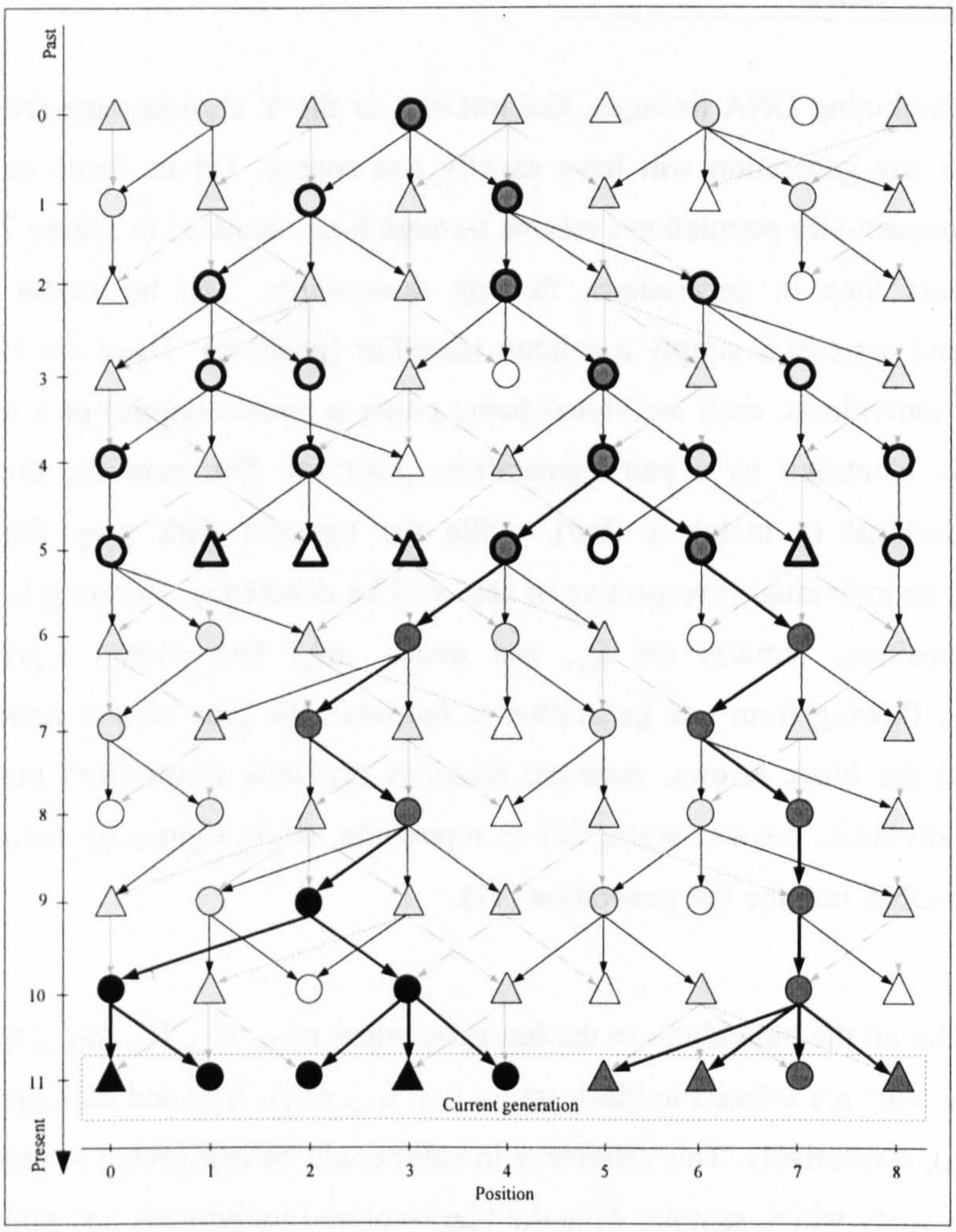

Figure 75: Example of *mtDNA* genealogy.

Vertical axis: time (generations), flowing downwards. Horizontal axis: non-directional, simply a column identifier (position). There are 12 generations, each with 9 individuals (females = circles or males = triangles). Arrows = parental relationships; gray arrows = paternal relations; black arrows = maternal relations (mtDNA flow). White individuals fail to reproduce. Light gray individuals fail to contribute mtDNA into the last generation. Dark gray and black individuals either belong to the last generation or contributed mtDNA into it (see text for details), while the heavy bordered individuals either belong to generation 5 or contributed mtDNA into it.

Up one generation, the mtDNA contained within the three females $f_{10,0}$, $f_{10,3}$ and $f_{10,7}$, traces back to only two females in generation 9, namely $f_{9,2}$ and $f_{9,7}$; all the other individuals in this generation failed to transmit their mtDNA into the present either because of reproductive failure ($f_{9,6}$), because they are males ($m_{9,0}$, $m_{9,3}$, $m_{9,4}$ and $m_{9,8}$) or because, even if being

females and producing viable offspring, all this children proved to be males ($f_{9,5}$), or, in the case of $f_{9,1}$, the only daughter did not reproduce. Thus, transmitting one's mtDNA into the next generation can be tricky, but it is even more so when just two generations are considered, as one has to have daughters which, in turn, have to have children. If one is a healthy female with an incredible fitness[301] of 10, but all these children are males, then her mtDNA will not be represented in two generations' time, but her nuclear genes will.

In generations 8, 7, 6 and 5, the number of distinct mtDNA lineages surviving into the last generation (11) remains 2, but, finally, they originate from the same female ($f_{4,5}$) in generation 4; $f_{4,5}$'s mtDNA, in turn, originates from her mother ($f_{3,5}$) and her grandmother ($f_{2,4}$) and so on ($f_{1,4}$ and finally $f_{0,3}$). Thus, all the mtDNA in the last generation originates from *a single* female, $f_{4,5}$, living seven generations into the past: she surely is the *mitochondrial Eve* of our toy world! As we turn back history, each time two or more separate lineages converge because a single female has more than a single offspring, a *coalescence event* takes place. For example, in generation 10, three coalescence events happened: two mitochondrial lineages carried by $m_{11,0}$ and $f_{11,1}$ coalesced into $f_{10,0}$, three mitochondrial lineages carried by $f_{11,2}$, $m_{11,3}$ and $f_{11,4}$ coalesced into $f_{10,3}$, and, finally, four lineages carried by $m_{11,5}$, $m_{11,6}$, $f_{11,7}$ and $m_{11,8}$ coalesced into $f_{10,7}$. There is another coalescence happening in generation 9 (the lineages of $f_{10,0}$ and $f_{10,3}$ coalesce into $f_{9,2}$), and a final one in generation 4 (the lineages of $f_{5,4}$ and $f_{5,6}$ coalesce into $f_{4,5}$). When two lineages coalesce, the individual into which this coalescence happens is their *most recent common ancestor* or *MRCA*. For example, the MRCA of $m_{11,0}$ and $f_{11,1}$ is $f_{10,0}$, the MRCA of $m_{11,0}$ and $f_{11,2}$ is $f_{9,2}$, while the MRCA of $m_{11,0}$ and $m_{11,8}$ is $f_{4,5}$. Thus, the MRCA of all mtDNA lineages in generation 11 lived a mere 7 generations before, and only 5 coalescence events have occurred. Moreover, even if all the maternal ancestors of $f_{4,5}$ are implicitly also common ancestors of all the mtDNA lineages in generation 11, there is *only one* most recent common ancestor of these lineages.

The existence of the MRCA of a set of mtDNAs is a logical necessity, given that any lineage must have a single direct ancestor, but can have any number of direct descendants. In this context, the existence of the mitochondrial Eve is not unexpected, nor a real discovery. When looking forward in time, it is impossible to predict which one of the living females is

301 The number of offspring reaching reproductive maturity; this represents a gross simplification of a very complex reality, but fits our needs (Skelton, 1993:165-166).

going to be a mitochondrial Eve, nor after how many generations. For example, consider $f_{4,0}$, a contemporary of our MRCA, $f_{4,5}$: why isn't she the mitochondrial Eve of our world? Simply because her mtDNA line was broken in generation 9 by her grand-grand-grand-daughter $f_{8,1}$, which had only one son, $m_{9,0}$, which couldn't transmit his mtDNA to his own daughter, $f_{10,0}$, who, in turn, inherited her mtDNA from her mother ($f_{9,2}$, a descendant of $f_{4,5}$ and bearer of this mtDNA lineage). It must be stressed, though, that other of the $f_{4,5}$'s genes are still present in the last generation, as there is a continuous line of descent connecting her to $m_{11,0}$, $f_{11,1}$, $f_{11,2}$, $m_{11,3}$ and $f_{11,4}$, which do carry (part of) her nuclear genes. This represents a clear case of decoupling between the histories of various genes and cautions against hasty generalizations based on only a handful of genes. It is also important to highlight that even if $f_{4,5}$'s mtDNA is inherited by all living individuals, it is highly probable that other sequences of her DNA have been lost.

Moreover, if we focus on generation 5 (the heavy-bordered individuals in Figure 75), its mitochondrial Eve is represented by $f_{0,3}$. This female is, necessarily, also a common ancestor of all the mtDNA in generation 11, because, by being the MRCA of all mtDNA lineages in a previous generation (5 in our case), she is also the ancestor of the MRCA of all these individuals ($f_{4,5}$ in out case). Thus, the mitochondrial Eve depends on the specific population considered, both spatially (geographic population or sampling procedure) and temporally (when did the composing individuals live)[302]. It must be noted that the individuals composing the population must not necessarily be contemporaneous, and this observation allows us to consider the MRCA of living and fossil humans.

Generally, let us restrict the concept of coalescence to only two lineages (Halliburton, 2004:455-456; Relethford, 2001:83). A *gene genealogy* represents the lines of ancestry connecting a set of lineages to their MRCA (Halliburton, 2004:456; Jobling, Hurles & Tyler-Smith, 2004:183). For a sample of *n* lineages, *n-1* coalescences connects them to their MRCA (Halliburton, 2004:456); in our case, 8 binary coalescences are enough (by conceptually splitting a triple or quadruple coalescence into two binary coalescences). It can be shown (Halliburton, 2004:456-458) that, in the case of mtDNA, the *expected coalescence*

302A very intuitive illustration is offered by Barbujani *et al.* (1998:489): "[...] suppose that some Europeans colonize Mars next year: If they successfully establish a population, the common mitochondrial ancestor of their descendants will be Palaeolithic. But it would not be wise for a population geneticist of the future to infer from that a Palaeolithic colonization of Mars."

time in generations (from *n* lineages to their MRCA), is given by:

$$E_{CoalescenceTime}(n, N_f) = 2N_f(n-1)/n$$

where N_f represents the *female effective population size* (see Section 2.2.3), but, because this is the *expected* value of a random variable, a *95% confidence interval* (CI) is usually given. In our example, because the population size is constant and the sex ratio is 0.5, N_f is approximately 4.5 and, thus, $E_{CoalescenceTime}(9, 4.5) = 8$ generations.

Annex 2. Politics and human evolution

Politics is a very distorting force when applied to science. A classic and well-known example (mainly because it happened so long ago and in a different socio-political context) is represented by the repression of genetics by Trofim Lysenko in the USSR during the period 1930s-1960s, purely on ideological grounds (Sheehan, 1993; Davies, 1997). But examples of "Lysenkoism" abound, many not so disastrous or obvious to the naked eye. One such case seems to be the modern political correctness as applied, implicitly or explicitly, to human evolutionary theorizing, especially the issue of *racism*. The best treatment to date of this issue seems to be Wolpoff & Caspari (1997).

As discussed, for example, in Banton (1998), racism is a body of attitudes and justifications having as their main effect the discrimination of humans on the basis of group characteristics, irrespective of their gender or religion. Racism is a fuzzy concept, extremely hard to define, but, it seems to focus on biological differences between groups. Banton (1998) describes a series of such concepts, viewed historically, but not necessarily replacing each other in succession: designation, lineage, type, subspecies, status, class and social construct. Racism is emphatically rejected by most modern scientists, especially those working in human evolution, probably as they have a first-hand experience with the unity of humanity as well as its diversity (Wolpoff & Caspari, 1997).

Some try to link the justification of racism to a "scientific" theory of the origins of races and, because of this, human evolution is a predilect target for misinterpretations, simplifications and outright false attributions[303]. As discussed in Section 2.1.3, polygenism was frequently adduced as "scientific proof" for the purported European superiority, even when it was emphatically rejected on scientific grounds as an accurate description of the fossil record and modern human variation. Nevertheless, it generated what can be called a "race-scare" which persists in the minds of scientists like a burning memory of disgrace and shame. But it seems that now this race-scare is used as an argument capable of deciding between competing theories, which is one of the worst distortions of the scientific process imaginable.

303 Another contributing factor is represented by the field's own history (see Wolpoff & Caspari, 1997).

More exactly, because science is a public process, some issues are perceived as important for a large sector of the public (like the issue of racism) and as large sectors of the public (weighted by other socio-economic factors) are important for the political process, arguments external to science are brought to bear on scientific disputes. In turn, as described in Powledge (2005) for the *Homo floresiensis* case, science is done by people, with their unavoidable range of biases, agendas and interests, and there is a strong pressure on media coverage as this, more often than not, brings funding and celebrity. She says:

> [t]he Hobbit [*Homo floresiensis*] story seems *designed for twenty-first century media because indeed it was*. [...] The Hobbit tale is a natural draw, featured in print and broadcast media everywhere. [...] Formerly inconspicuous palaeoanthropologists, anthropologists and microcephaly experts are suddenly in demand (Powledge, 2005:611, *italics* mine),

and, as always when the media intervenes and excessive popularizations, simplifications and metaphors start flowing, the scientific process is caricatured as a quarrel between opposing personalities and puny interests and agendas, the public can get only one message: science is a mess, no truth can be found because everybody argues with everybody else and previous "truths" are overturned by current ones: it's a madhouse[304]. As Powledge says (citing John Hawks), in the case of the Flores man:

> [...] the dispute has been bad for palaeoanthropology and good for creationism. Searching the World Wide Web for information on the Hobbit, [...] uncovers many creationist sites [...] [saying] 'Look! These people don't know what they're doing! They don't know what they're talking about! They're disagreeing about the most basic issues – about whether something is diseased or not!' [...] Given the amount of media attention, it just makes the field look incompetent (Powledge, 2005:611).

I would say that what was bad for the field (because it certainly was, given the current rise of Creationism, post-modernist criticisms of science and other such inept doctrines – Gross & Levitt, 1998; Dawkins, 2004), is *not* that there is disagreement, even if sometimes the arguments became personal[305] but that such disagreements are distorted into trivial media shows. Another recent and very disturbing example is represented by the hijacking of two recent publications in *Science* (September 2005, Mekel-Bobrov *et al.*, 2005 and Evans *et al.*, 2005), reporting the study of two microcephalia-related human genes (*ASPM*, OMIM 605481 and *Microcephalin*, OMIM 607117; Chapters 4 and 5) and finding that these genes present each a derived haplogroup which show signatures of natural selection and strong

304This fits very well in the "relativistic" state of mind described by Gross & Levitt (1998).
305The same things happen also, for example, in evolutionary biology – see Dawkins vs. Gould.

spatial structure – and their capitalization by individuals of extreme right-wing orientations and used to justify racist attitudes[306].

It is the normal way of science to nurture controversies: this is the way ideas are tested and adapted to reality. A field of study where there are no controversies is either dead, or plainly non-scientific, dogmatic. That human evolution in particular is home to such disputes must be taken to mean only one thing: it is still vigorous, still young, still promising. The problem, again, is not the amount of media attention, which, in principle at least, should do no harm, but the *way* media pitches the controversies. And what would make the field look *competent* in the modern media coverage, anyway? A bunch of people dictating the one and only truth revealed by the bones/gods/gurus? I think the solution is to present science the way it *is* done but making clear *how* it does things, not the aseptic, artificial, linear fantasies disguised as "popularization of science"[307].

Milford Wolpoff and Rachel Caspari recount (from their own point of view) the beginnings of the political immixture into the current scientific debate concerning modern human origins: in a 1988 paper in *Natural History*, shortly after the publication of the original Eve theory (Cann, Stoneking & Wilson, 1987) and related proposal of ROA (Stringer & Andrews, 1998), the evolutionary biology popularizer Stephen Jay Gould[308] declaims:

> [a]ll modern humans form an entity united by physical bonds of descent from a recent African root; we are not merely the current state of a tendency as the multiregional model suggests. Our unities are genealogical; we are an object of history (Gould, 1988, cited in Wolpoff & Caspari, 1997:54).

This was shocking, because

> [b]y appealing to the implication that it demonstrated we are all brothers under the skin, the unspoken but implicit charge is that the opposing view (ours [multiregionalism]) somehow shows we are *not* 'brothers under the skin'. In the Natural History article Gould, for the first time, placed the debate in the arena of political correctness, and political *in*-correctness was clearly attributed to our side

306See for example the controversial Steve Sailer's posting (http://www.vdare.com/sailer/050911_new_orleans.htm) where allusions to *ASPM* & *Microcephalin* are immersed into a racist pleading against African Americans, which grossly confuse social and biological issues.

307See for example Dawkin's discussion of such a BBC documentary in Dawkins (2004:57-59).

308Controversial in himself, seen by many non-biologists as *the expert* in evolutionary biology but criticized by many leading figures of the field for his unorthodox and distorting views (e.g., Richard Dawkins and John Maynard-Smith). Especially telling from this point of view is his promoting of punctuated equilibrium (Eldredge & Gould, 1972) to the rank of a revolution in evolutionary biology.

of it (Wolpoff & Caspari, 1997:54, *italics* in original).

This at least questionable move out of science and into the *circus* was to mark almost 20[309] years of research on modern human origins, with the popular image that ROA *must* be true because all the alternatives *must* be politically incorrect, thus *wrong*. The sloppy logic and confusion of domains[310] did not bother the public, and even scientists endorsed it or became biased by it (Wolpoff & Caspari, 1997). For example, Stringer & McKie claim that:

> [s]uch a theory [multiregionalism] would suggest, at face value, that modern humanity's constituent races are divided by 'fundamental and deep-rooted differences (Stringer & McKie, 1996:49; but see pages 48-50 for the entire discussion),

while the copiously wrong assertions by York (2005):

> [m]ultiregionalists adhere to the position that the division of humans into distinct groups (races) is very old, which implies that genuine biological differences exist among contemporary races. [...] It is important to note in all fairness that contemporary supporters of multiregionalism typically deny any support for racist views or policies and acknowledge the high level of genetic similarity among human populations, but the multiregionalist position does, nonetheless, reify divisions of humans into distinct biological races (if not species) (York, 2005)

can only prove the level of political content of uninformed opinions and the moral judgments masquerading as scientific criteria (is this but an example of a global trend? see Gross & Levitt, 1998).

Of course, political and moral arguments *cannot* be used now to "redeem" multiregionalism in an act of moral reparation of some sort: it would be exactly the same failure of the scientific method. Moral and political arguments have *no value* in a scientific dispute, and the dream of a socially- and politically-involved scientist is, to be mild, far worse than simply wrong, even when clad in nice words as "progressism"[311] and the like:

> [o]f course, anthropological work has also been used to support progressive social causes by both the scientists themselves and political agencies using the results of their work. But one *generation's progressive social causes can become the repressive policy of the next*, as the history of the eugenics movement so clearly

309I dare to see signs of this immixture fading away, as more an more people become aware of it.

310This is so ironic, to have originated from the even creator of NOMA (Non-Overlapping Magisteria, Gould, 1987), designed exactly in order to clarify such a confusion of domains (religion vs. science).

311As someone living a good part of his childhood and youth in a "communist Eden of equality and progressism", I think I know first hand how important is to be able to think and act in a politically free environment. I do think we don't need new Lysenkos, no matter what flag they fight for and what dreams they try to impose upon us.

shows (Wolpoff & Caspari, 1997:11, *italics* mine).

I want to highlight again that I do not advocate for multiregionalism or another admixture model and against ROA on political and moral grounds. But it is hard not to observe the power of misrepresentation and simplification in forming strong, widespread opinions: it is almost a wonder to try to understand how a human evolutionary model based on a global and continuous network of gene flow can be caricatured as racist. How can it be that a model in which the even question concerning the "origin of races" is meaningless, all populations having a long history of admixture from all over the world, to varying degrees, is pictured to claim distinct origins of races? Taking the things "at face value", in such a model, no one can claim in any meaningful way to belong to a separate race, while, given that ROA is a candelabra model, albeit a fairly recent one (Templeton, 1998), it is exactly in this supposedly progressist and politically correct model that races are viewed as distinct lineages[312]. That this step is very easy to make is shown, for example, by Vincent Sarich and Frank Miele's book (Sarich & Miele, 2004), where superficial scientific arguments (ROA is fundamental) and an incredible lack of historical knowledge and understanding, doubled by such a parochial world view that even the legendary Middle Ages village (Davies, 1997) would shine as a beacon of intellectual openness, are used to argue that racism, and racial hierarchies are "real"[313].

Without any desire to blame anyone in particular, all this being probably an indirect effect of the love for simplicity in an unknown domain of knowledge, it still seems probable that Gould's own incline towards seeing discontinuities (punctuated equilibria) and speciations, combined with his political convictions, offered the starting point in this sterile direction[314].

312Even implicitly, in the usage of trees to depict evolutionary relationships between populations (Wolpoff & Caspari, 2000).

313This book is a must read, as one of the best example of pure racist nonsense: all the "classical" arguments are marshaled, including the IQ inter-"racial" differences (R. Lynn is copiously cited), but no understanding of the other aspects of IQ is shown. It is incredible that such a book can still be published in 2004.

314Gould's misunderstanding of multiregionalism is well documented (e.g., Gould, 2002).

Annex 3: How bad can it get? Language-genes correlations with an agenda

The dangers of bad interdisciplinary research are many, especially when unfamiliar but fashionable concepts from unfamiliar but fashionable fields are used, and the results can be seen either as hilarious or disastrous. I have selected as an example Arnaiz-Villena, Martińez-Laso & Alonso-Garciá (2001) as it has the rank of 1 on a Google[315] search for "correlation languages genes" and, so it is highly visible for anyone interested in the subject[316].

The authors collected frequency data on the HLA system as the genetic side of the study (see below). In what concerns the linguistic side, they state that:

> Once shown, the contradictory (and fruitless) current dogma [the comparative method?] for approaching decipherment, we have followed a methodology which is similar to that proposed by Greenberg and Ruhlen [...] (Arnaiz-Villena, Martińez-Laso & Alonso-Garciá, 2001:1053).

Their "premises" are:

1. Languages may correctly be classified and decipherment approached with 10-20 "diagnostic" cognates [...]
2. Most of the written ancient Mediterranean languages studied previously by us (i.e., Iberian-Tartesian, Etruscan, Linear A, etc.) refer to an apparently common religion [...]
3. Most of these deciphered "*Usko*-Mediterranean" languages refer to the following matters: Religion and after death [...] [and] Accountancy related to food-storage and other topics [...]
4. There are groups of words that are found together in the different languages [...]
5. Beginning and ending of words are problematic and unless meaning is known, it is very difficult to separate them [...]
6. Common and proper names are almost impossible to distinguish [...]
7. Basque language has remained with little modifications through time, because invasions have not modified this and other Basque society characteristics [...]
8. Basque language was much more extended that its present day limits [...]

(Arnaiz-Villena, Martińez-Laso & Alonso-Garciá, 2001:1053-1054, *italics* in original).

315www.google.co.uk. The search was done in September 2006.

316For example, Cavalli-Sforza's "Genes, Peoples and Languages" ranked only 6 and 7 on the same search.

Their Table 1 (Arnaiz-Villena, Martińez-Laso & Alonso-Garciá, 2001:1055), containing the list of "words that are found together in the different languages" (their premise 4) is hilarious to the extreme. It contains 21 English "meanings" (with Spanish equivalents) and their rendition in the 11 "*Usko*-Mediterranean" languages "deciphered" by the authors: Basque, Iberian-Tartesian, Etruscan, Minoan, Berber, Punic-Carthaginian, Hittite, Sumerian, Eblaic, Elamite and Egyptian. Most of the words are monosyllabic (with some bisyllabic) and a representative example is offered by the first entry, glossed in English as "Father, Panel, Cleft", Spanish "Padre, Panel, Hendidura", Basque "Aba", Iberian-Tartesian "Aba", Etruscan "Ava", Minoan "Aba", Berber "Aba", Punic-Carthaginian "Aba", Hittite "Aba", Sumerian "Aba", Eblaic "Aba", Elamite "Aba" and Egyptian "Aba". Unfortunately, they do not give the actual meaning of this monosyllabic word in each of the concerned languages, and the English gloss does not seem extremely coherent (what connection could there be between father, panel and cleft?). Also, unfortunately for the authors' thesis, a quick search of readily available online sources shows that, for example, while "aba" does exist in Basque, meaning "father" (in the religious sense), is a neologism[317], for Etruscan the closest match seems to be "apa" (father)[318] and Sumerian has an "ab (abba)/ab-ba/abba$_2$" meaning "old (person); witness; father; elder; an official"[319], while the Hittite "aba" turns out to be the Akkadogram "*ABA*" (father) and not a Hittite word at all (Güterbok & Hoffner, 1997:217, 299)[320]; and the other proposed "diagnostic" cognates do not fare better. Moreover, it is very well known that mono- and bi-syllabic words are not, in general, acceptable for proving genetic relationships, as the probability of coincidence is far too high[321]. And, besides, this amalgamation of linguistic isolates (e.g., Basque, Etruscan, Sumerian) with Afro-Asiatic (Punic-Carthaginian, Berber, Egyptian) and Indo-European (Hittite) seems to hint at the authors' lack of understanding of what cognation is: it certainly is *not* (near)identity of form for (very) loosely corresponding meanings. The explanations offered by the authors in the "Translation and transliteration" section of the paper (Arnaiz-Villena, Martińez-Laso & Alonso-Garciá, 2001:1054) adds to the feeling of transgressing science:

> Berber has been distinguished from the Arab *contamination* by comparison with Basque, Iberian-Tartesian, and Arab (Arnaiz-Villena, Martińez-Laso & Alonso-Garciá, 2001:1054, *italics* mine).

317For example http://www1.euskadi.net/hizt_el/eusk.asp?Sarrera=aba http://www1.euskadi.net/harluxet/hiztegia1.asp?sarrera=aba1 or http://www1.euskadi.net/hizt_el/eusk.asp?Sarrera=aba (September, 2006).
318http://etruskisch.de/pgs/vc.htm, September, 2006
319http://psd.museum.upenn.edu/epsd/epsd/e72.html, September, 2006
320The actual Hittite word for father is "attaš" (Güterbok & Hoffner, 1997:12).
321See also Jakobson 's (1971) *sound symbolism*. Thanks to J. Hurford for comments.

One is left to wonder why not comparing it also with Martian? This would have certainly helped the authors get rid of all those nasty Afro-Asiatic "contaminations".

The results of the study, as expected, are... intriguing, for lack of a better word. From a purely genetic point of view, the populations phenogram obtained from HLA frequency data seems quite standard, except for two major glitches: San, Japanese, Egyptians and Italians form a subclade, while Greeks cluster with sub-Saharan Africa instead of their European and Near-Eastern neighbors (Arnaiz-Villena, Martiñez-Laso & Alonso-Garciá, 2001:1054, 1056). The first glitch is simply glossed over (really hard to explain without invoking Martians translocating people across the world in flying saucers), but for the second, the authors, with their already familiar obstinacy, instead of questioning the quality of their genetic data, go on and construct a story containing a

> [...] migration from southern Sahara which mixed with ancient Greeks to give rise to a part of the (normal case) genetic background. The admixture must have occurred in the Aegean Islands and Athens area at least [...] Also, the time when admixture occurred could be after of some of the Negroid Egyptian dynasties (Nubian or from other periods) or after undetermined natural catastrophes (i.e., dryness) (Arnaiz-Villena, Martiñez-Laso & Alonso-Garciá, 2001:1056).

I still want to hold to my belief that Martians are actually the best explanation for this. Nevertheless, a much simpler (and earthly) alternative is offered by the fact that the HLA (human leukocyte antigen) system is functionally involved in the immune system (Seeley, Stephens & Tate, 2006:796-820; Jobling, Hurles & Tyler-Smith, 2004:139-140) and thus, one of the eminent non-neutral genetic systems (Cavalli-Sforza, Menozzi & Piazza, 1994:131, 142; Jobling, Hurles & Tyler-Smith, 2004:139-140).

But the peak of arrogance towards historical linguistics, of which the authors don't seem to grasp much, is reached in their "The Usko-Mediterranean languages" section, in which they build the case of a circum-Mediterranean ancient language family. To make things fit into this agenda, after invoking migrations out of the post-LGM drying Sahara[322] and forcing many scarcely known languages to be related in their peculiar way, they still face one major obstacle: Hittite *is* an Indo-European language. But wait:

> Hittite was classified by Hrozny [Hrozny, 1915] as Indo-European with the study of only one phrase, which is now translated by us with the help of the Basque-Spanish equivalences:

322The timing of this drying seems to not quite well fit this scenario (Mithen, 2003; Wilson, Drury & Chapman, 2000)

HITTITE: NU NINDA-AN EZZTENI VADAR-MA EKUTTENI (Full Text)

Basque: NUN_INDA_N_EZ_Z(U)

Spanish: Donde-En el pantano-No-Fuego

English: Where-In the bog-Not-Fire

Basque: ATE-NI-BA-TAR-(A)MA-EKUTE-NI

Spanish: La puerta-Yo-Si-Procedente-La madre-Pertenezco-Yo

English: The door-I-Yes-Coming from-The mother-belong-I

Full English translation: "Where in the bog (is) not fire, yes I (am) coming from The Door, I belong to The Mother" (Arnaiz-Villena, Martińez-Laso & Alonso-Garciá, 2001:1058),

which translation, by the way, makes a terrible sense (it almost sounds like true Martian).

The history of the discovery of Hittite and its Indo-European affiliation is much more complex and interesting (Fortson, 2004:154), but the authors seem to have a very partial knowledge of it. The actual phrase is (for example, in a larger context, gray background; Güterbok & Hoffner, 1997:6-7):

nu LUGAL-*aš udd*[*ā*]*r*=*mit* [*pa-aḫ-ḫa-aš-d*]*u-ma-at nu* NINDA-*an azzašteni wātarr*=*a ekutteni*... *mān AWAT* LUGAL=*ma UL paḫḫašnutteni*

and its translation is (*italics*) "[Obs]erve my, the king's, words. Then *you will eat bread and drink water*... But if you do not observe the king's words (you will not stay alive)." Capital letters stand for Sumerograms, like "NINDA" ("food, bread"), from which Hrozný's translation started, and italic capital letters stand for Akkadograms, like "*UL*" ("not")[323]. But even *if* Hrozný's translation would have been wrong, the existence of the Anatolian languages, comprising Hittite, Luvian (Cuneiform and Hieroglyphic), Palaic, Lycian, Lydian, Carian, Pisidian and Sidetic, and their certain affiliation to the Indo-European family[324], rests on an impressive corpus of historical linguistic work of the highest quality. After Hrozný's seminal paper, history did not stop, as the authors seem to think, but an impressive corpus of texts[325] have been unearthed and our current linguistic reconstruction of Hittite is quite coherent (Fortson, 2004:154-177; Güterbok & Hoffner, 1997). What the authors completely ignore is that these Anatolian languages are not some fancy new branch grown into the Indo-European tree based on a handful of sketchy "diagnostic cognates", but

323For a description of these conventions, see Fortson, 2005:160-161.

324Either as a sub-family or as a sister branch.

325E.g., the corpus of Hittite texts, maintained and updated by Dr. B. J. Collins, Department of Near Eastern Studies, Emory University, Atlanta.

that they represent a very conservative branch and their phonology confirmed a prediction made by Ferdinand de Saussure half a century before their discovery concerning the existence of laryngeals in PIE (Fortson, 2004:75-76), of which Anatolian languages alone in the entire Indo-European family retained certain direct reflexes (e.g. the velar fricative ḫ in Hittite and Luvian) (Fortson, 2004:56).

Thus, this paper proves a total lack of knowledge and respect towards linguistics and, even from a purely genetic point of view, its methodology is seriously flawed. It seems to fit perfectly the words of the regretted historical linguist and specialist in Basque linguistics, Larry Trask:

> But please note: I do not want to hear about the following:
>
> - Your latest proof that Basque is related to Iberian / Etruscan / Pictish / Sumerian / Minoan / Tibetan / Isthmus Zapotec / Martian;
>
> - Your discovery that Basque is the secret key to understanding the Ogam inscriptions / the Phaistos disc / the Easter Island carvings / the Egyptian Book of the Dead / the Qabbala / the prophecies of Nostradamus / your PC manual / the movements of the New York Stock Exchange;
>
> - Your belief that Basque is the ancestral language of all humankind / a remnant of the speech of lost Atlantis / the language of the vanished civilization of Antarctica / evidence of visitors from Proxima Centauri.
>
> I definitely do not want to hear about these scholarly breakthroughs[326]

And yet, despite all these, this paper was published by a high-ranking (2.7 impact factor) peer-reviewed journal[327]: this can be in large part explained, I think, by the non-appropriateness of the paper's content for the journal's areas of expertise. Therefore, the reviewers cannot be expected to have been experts in these subject matters, and the overall tone of the paper, excluding its actual contents, seems convincing and authoritative.

The first main conclusion to be drawn from the analysis of this case seems to be that such scientifically flawed studies can potentially do much damage to the field, by exposing newcomers to a barrage of distorted information, incorrect methodology and overt despising of legitimate knowledge and opinions when they do not fit the desired picture, everything disguised as "scientific", and by creating an impression of amateurish babbling, extending

326From his Basque WEB page, as currently hosted here: http://www.buber.net/Basque/Euskara/Larry/WebSite/basque.html (an archive from 1996). September, 2006.

327"Human Immunology", homepage http://www.elsevier.com/wps/find/journaldescription.cws_home/505763/description, September, 2006

over the entire field. The second main conclusion concerns the limits of expertise and their transgression: while there is no reason to doubt the authors' high competence in their respective domains, they seems utterly unable to realize that stepping outside their fields is not warranted.

Annex 4: Nettle & Harriss (2003) revisited

As discussed in Section 3.2.4.6, the method applied by Nettle & Harriss (2003) to the study of genes-languages correlations is interesting but has some potential problems. Therefore, I have decided to try to adapt and apply it to the present data (Section 4). Nettle & Harriss (2003:333) have divided their large sample (102 populations) into 5 regions: Europe (25), West Asia (18), East and Central Asia (21), Southeast Asia (24) and West Africa (13). Given the reduced sample size (49) available in this study and its different geographical distribution, a slightly different division scheme was used here: Europe (8: FrBasque, French, Sardinian, NItalian, Tuscan, Orcadian, Russian, Adygei), West Asia (3: Druze, Palestinian, Bedouin), East and Central Asia (20: Hazara, Balochi, Pathan, Burusho, Makrani, Brahui, Kalash, Sindhi, Hezhen, Mongola, Daur, Orogen, Miaozu, Yizu, Tujia, Han, Xibo, Uygur, Japanese, Yakut), Southeast Asia (7: Dai, Lahu, She, Naxi, Tu, Cambodian, NANMelanesian) and Africa (11: SESWBantu, San, Mbuti, Turu, Kikuyu, Biaka, Bakola, Bamoun, Yoruba, Mandenka, Mozabite).

The resulting correlations are in Table 43 (both Pearson's – as in the original Nettle & Harriss, 2003 – and Mantel's[328]). It must be highlighted that, in our case, each language is represented by a single population, and the inclusion of same-population distances (as in Nettle & Harriss, 2003[329]) would have artificially increased the correlations and linear regression fit. Thus, same-language pairs were excluded from the present study.

As expected on theoretical grounds (Bonnet & Van de Peer, 2002:2), while the two estimations of the correlation coefficient, Pearson's and Mantel's, are identical, the *p*-values of Pearson's estimate are much more liberal than Mantel's. For our data, in most cases, both Pearson's and Mantel *p*-values agree on the significance of the correlation at the 0.05 level, but there is one exception: East & Central Asia. Thus, supporting theoretical approaches (Bonnet & Van de Peer, 2002; Fortin & Dale, 2005; Mantel, 1967), using Pearson's *p*-value estimates for distance matrices is not generally warranted.

328Holm mcc was used to adjust the *p*-values.

329It is not clear in Nettle & Harris' (2003) sample for how many languages there are more than one populations speaking it, so that the impact of considering same-language pairs on their statistics cannot be evaluated.

Concerning the correlations between the *N-HLD* and genetic versus log genetic distances, even if the correlations with the log genetic distances are higher, there are no important differences between the two (confirmed by a paired-samples t-test, t = -1.8128, df = 4, p = 0.1441). The only notable difference is represented by the global case (r = 0.2771 vs r_{log} = 0.4553). It can be concluded, thus, that even if taking the logarithm of the genetic distances (as in Nettle & Harriss, 2003:334) slightly increases the linear correlations with *N-HLD*, this does not seem absolutely necessary.

Globally, all the correlations with *N-HLD* are significant, while regionally, *N-HLD* is correlated with geography only in East and Central Asia, and with genetics in Africa. The present data does not confirm Nettle & Harriss' (2003:334-335) results, but this is probably due to the highly subjective nature of regional classification of the populations, hypothesis which seems supported by the high heterogeneity of the correlations across regions.

The next step in Nettle & Harriss (2003:334-335) was to perform a linear regression of the log genetic distance on the geographic distances, for each region separately (Table 44).

Region	N-HLD vs geography			N-HLD vs genetics			N-HLD vs Log(genetics)			N-HLD vs linguistic features		
	r	$p_{Pearson}$	p_{Mantel}	r	$p_{Pearson}$	p_{Mantel}	r	$p_{Pearson}$	p_{Mantel}	r	$p_{Pearson}$	p_{Mantel}
Global	**0.382**	**0.0000**	**0.0005**	**0.277**	**0.0000**	**0.0005**	**0.455**	**0.0000**	**0.0005**	**0.445**	**0.0000**	**0.0005**
Europe	0.497	0.0216	0.1899	-0.286	0.2816	1.0000	-0.259	0.3680	1.0000	**0.873**	**0.0000**	**0.0081**
West Asia[330]	–	–	–	–	–	–	–	–	–	–	–	–
East & Central Asia	**0.519**	**0.0000**	**0.0005**	***0.193***	***0.0228***	***0.0705***	**0.216**	**0.0081**	**0.0477**	**0.384**	**0.0000**	**0.0005**
South-East Asia	0.208	0.3740	0.1899	-0.080	0.7304	1.0000	-0.103	0.6572	1.0000	0.144	0.5325	0.3345
Africa	0.181	0.3740	0.1914	**0.725**	**0.0000**	**0.0005**	**0.893**	**0.0000**	**0.0005**	**0.629**	**0.0000**	**0.0140**

Table 43: The correlations (Pearson & Mantel) between the *N-HLD* and geographic, genetic, log(genetic) and linguistic features distances.

Gray, bold: both Pearson & Mantel significant at the 0.05 level, light gray, italic bold: only Pearson is significant at the 0.05 level (Holm mcc).

Region	Intercept (A)			Geography (B)			Adjusted R^2	df	p-value
	Estimate	Std. error	p-value	Estimate	Std. error	p-value			
Global	**-3.6900**	**0.0482**	**0.0000**	**0.0001**	**0.0000**	**0.0000**	**0.132**	**1174**	**0.0000**
Europe	-3.3630	0.1832	0.0000	-0.0001	0.0001	0.2010	0.026	26	0.2006
West Asia	-4.4976	0.0234	0.0033	0.0018	0.0003	0.0923	0.958	1	0.0923
East & Central Asia	**-3.4680**	**0.0468**	**0.0000**	**0.0001**	**0.0000**	**0.0000**	**0.130**	**188**	**0.0000**
South-East Asia	**-2.8540**	**0.0547**	**0.0000**	**0.0001**	**0.0000**	**0.0003**	**0.475**	**19**	**0.0003**
Africa	-5.6945	0.9336	0.0000	0.0003	0.0003	0.2390	0.008	53	0.2394

Table 44: The linear regression of log(genetic distances) on land distances, as in Nettle & Harriss (2003:334-335).

330Only three data points.

The linear regression of *log*(genetic distance) on land distance is highly significant globally but explains a small fraction of the variance (adjusted $R^2 = 0.1322$), and is non-significant in Europe, West Asia and Africa[331], but is highly significant in East and Central Asia, where it also explains few of the variance (adjusted $R^2 = 0.1304$), and in South-East Asia, where it explains an important part of the variance (adjusted $R^2 = 0.4745$). These results do not concord with Nettle & Harris (2003:334-335) except in one respect: the intercept and regression coefficient estimates are comparable[332], as confirmed by a paired-samples t-test between Nettle & Harriss' (2003:335, Table 2) *A*'s and *B*'s and the absolute values of our *A*'s and *B*'s ($t = 1.3618$, $df = 9$, $p = 0.2064$, failing to reject the null hypothesis).

Next, the residuals of these regressions were computed and sorted by *N-HLD*, as in Nettle & Harriss (2003:335). ANOVA was performed for

residuals ~ *N-HLD*

for each region separately and the results are in Table 45 and Figure 76.

As opposed to Nettle & Harris (2003:335-338), the only significant ANOVAs were found in the global case and for Africa. Moreover, the boxplots[333] of residuals versus *N-HLD* do not show any clear trend except for Europe (residuals' median decrease with decreasing linguistic relatedness) and Africa (the reverse pattern). The only region showing a pattern like the one detected by Nettle & Harriss (2003:335-338) is Africa, but even here it is not easy to interpret, in the sense that with increasing linguistic dissimilarity, the genetic similarity increases relative to the expectancy based on geographic distance alone (the opposite of the neat pattern found in Europe by Nettle & Harris (2003:335 and Figure 1, p. 336).

331It must be pointed out that this is not due to the small number of data points, as Europe and Africa (non-significant) have $df = 26$ and 53, respectively, while South-East Asia (significant) has $df = 19$.

332Except for the intercept's sign, as in Nettle & Harriss (2003:335, Table 2) they are reported positive, when, due to the fact that the genetic distance they use (F_{ST}) is defined as taking values between 0 and 1 (e.g., Jobling, Hurles & Tyler-Smith, 2004:168), then the $log(F_{ST})$ would be expected to be negative, and given that the land distances are naturally positive, one would have expected the intercepts to be negative: $log(F_{ST}) = A + B*landdist$, and $landdist > 0$, $B > 0$, $A > 0$, would imply $log(F_{ST}) > 0$, which would imply in turn that $F_{ST} > 1$, which is impossible. Thus, I assume in the following that Nettle & Harriss (2003) reported the absolute value of the intercepts *A*.

333These are not directly comparable to the error bars in Nettle & Harriss (2003) – thanks to D. Nettle for the comment.

Region	N-HLD = 2				N-HLD = 3				N-HLD = 4				ANOVA p-value
	No.[334]	Min	Mean	Max	No.	Min	Mean	Max	No.	Min	Mean	Max	
Global	**47**	**-9.660**	**-1.475**	**1.069**	**92**	**-1.014**	**-0.143**	**0.699**	**1037**	**-0.854**	**0.080**	**1.478**	**0.0000**
Europe	6	-0.710	0.080	0.623	9	-0.549	0.037	0.769	13	-0.690	-0.062	0.714	0.4801
West Asia	3	-0.017	0.000	0.017	0	-	-	-	0	-	-	-	-
East & Central Asia	21	-0.753	-0.124	0.562	18	-0.207	0.225	0.440	151	-1.016	-0.010	0.592	0.7060
South-East Asia	1	0.252	0.252	0.252	0				20	-0.281	-0.013	0.270	0.1157
Africa	**10**	**-8.321**	**-6.179**	**-5.026**	**18**	**-0.496**	**0.629**	**1.620**	**27**	**-0.064**	**1.870**	**3.214**	**0.0000**

Table 45: The residuals versus *N-HLD* for each region.

The ANOVA p-value column reports the probability that there is a statistically significant difference between N-HLD distance classes. Gray bold: ANOVA significant at the 0.01 level.

334The number of pairs falling in this *N-HLD* distance class.

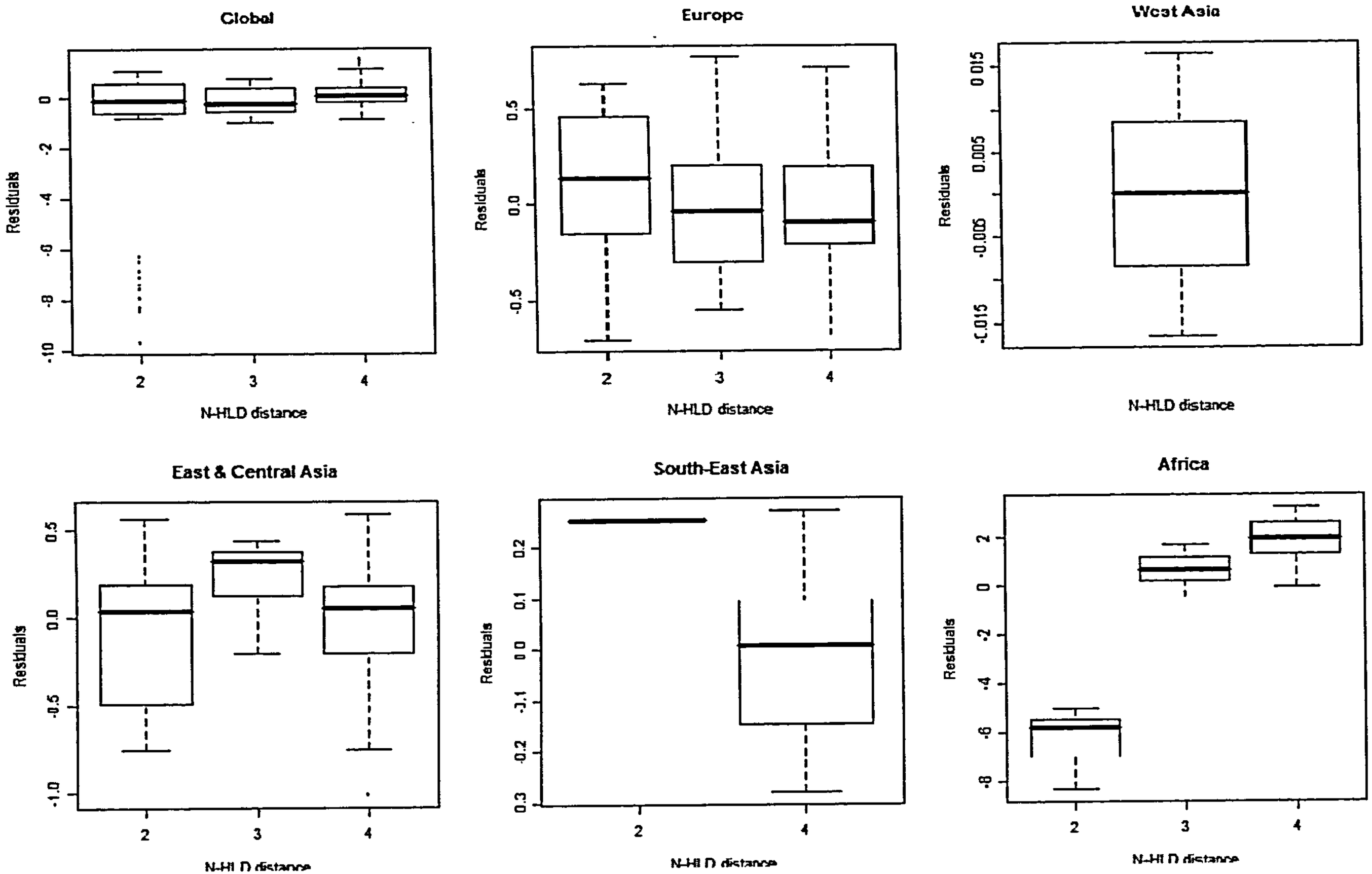

Figure 76: The boxplots of residuals vs. N-HLD for each region separately.

Overall, this kind of analysis seems to detect some patterning on a global scale, but at the regional level no consistent pattern seems to be discernible. Moreover, it has many inconsistencies and problems, and, therefore, it seems that this method is not entirely appropriate for geographical studies of genetic and linguistic relationships. It is possible that a better classification of the sample populations into regions would have helped the method, but there is no clear definition of what this "better" classification would be and if it would not inject the conclusions into the premises[335].

335For example, a definition of regions based on linguistic families or genetic similarity.

Annex 5: Description of the sample populations

In the following, for each of the 54 population in the *OWF sample*, a short description is provided, while later sections will detail the linguistic and genetic aspects[336]. Briefly, from a genetic point of view, two databases were used (ALFRED[337], Osier *et al.*, 2002, Rajeevan *et al.*, 2003, and HDGP, the Human Diversity Panel Genotypes) and some populations are contained only in one of them.

The following conventions have been used:

> **Population name** (*Population short name*, alternate names including other scripts; ~estimated population size). Short description, geo-political situation, language family (alternative language names [3-letter code]), geographical reference point (town, city, island or region). Samples corresponding to this population in ALFRED and/or HDGP databases.

For sub-Saharan Africa the genetic sampling is unexpectedly poor. Therefore, genetic data are (almost) completely missing for some populations, but both in ALFRED and HDGP there appears a very poorly specified composite sample named "Bantu speakers" (ALFRED PO000041F) and "Bantu" (HDGP), which seems to encompass various individuals belonging to Bantu-speaking groups across sub-Saharan Africa. This composite population was used to fill in the missing data in five sub-Saharan population (SESWBantu, Turu, Kikuyu, Bakola and Bamoun), so that these populations can be retained for analysis. This missing data handling procedure (Tabachnick & Fidell, 2001:58-66) was tested (see main text for details) and found to not distort the data too much.

Sub-Saharan Africa[338]: 14 populations belong to this geographical region, but most of them are unexpectedly poorly sampled from a genetic point of view. Also, the linguistic information concerning their languages is not readily available.

336The main sources of geographical, political, historical, genetic and linguistic information are: the Ethnologue (Gordon, 2005), the ALFRED database (Osier *et al.*, 2002; Rajeevan *et al.*, 2003), Wikipedia (www.wikipedia.org), Hourani (2002) for Arab-speaking populations, Davies (1997) for Europe and Cavalli-Sforza, Menozzi & Piazza (1994) for general information.

337The Allele Frequency Database, http://alfred.med.yale.edu/alfred/index.asp, September 2006.

338The splitting into macro-regions followed the HGDP database conventions.

Southeastern and Southwestern Bantu (*SESWBantu*; ~27,000,000). This obviously represents an amalgamated sample, containing speakers of Bantu languages reportedly living in South Africa, and probably including speakers of Ndebele (nbl), Sotho (nso, sot), Swati (ssw), Tsonga (tso), Xhosa (xho), Zulu (zul), among others (Gordon, 2005). The major city chosen was the South African capital, Pretoria. Linguistically, zul/xho have been used to represent this sample. Genetically, the ALFRED populations[339] PO000470L (Zulu) and PO000154K (Xhosa) were used. Also, ALFRED PO000041F ("Bantu speakers") and HDGP "Bantu" populations can be applied.

San (*San*; ~27,000). This sample is located in Namibia and probably includes hunter-gatherer groups living in the Kalahari desert and speakers of Ju|'hoan (ktz), Nama (naq), ǃXóõ (nmn) and/or Kxoe (xuu), among others (Gordon, 2005). Nama (naq) was chosen to represent this population linguistically, because of the number of its speakers and availability of information. The Namibian capital, Windhoek, was chosen as geographical reference. Genetically, ALFRED PO000073K ("Khoisan") and HDGP "San" populations were used.

Mbuti Pygmy (*Mbuti*; ~30,000-40,000). These groups of pygmy live in the Ituri forest of the DRC and speak a Nilo-Saharan language (Lese [les]), but due to information availability problems, the related Efe [efe] and Ma'di [mhi, smn] have been used. The geographical reference was taken to be represented by the capital of the Ituri province, Bunia. Genetically, ALFRED PO000006G and HDGP "Mbuti_Pygmies" populations were used.

Masai (~450,000). A poorly-sampled population speaking a Nilo-Saharan language (Maasai [mas]) and living in Kenya and Tanzania. Arusha, the capital of the region with the same name in Tanzania, was selected as geographical reference, and genetically, only the ALFRED database contained information about this population (PO000456P). Due to the lack of enough genetic information, this population was not considered in the following analyses.

Sandawe (~40,000). Also a very poorly sampled hunter-gatherer population, speaking a

339The ALFRED population UIDs given can be used to uniquely identify the population through the site's UID search engine: http://alfred.med.yale.edu/alfred/uidsearch.asp (September, 2006). Usually, the entries in ALFRED also contain short notes on the population's composition and history.

Khoisan language (Sandawe, [sad]) and living in Tanzania. The capital of Tanzania (the National Assembly) and of the Dodoma region, Dodoma, was chosen as the geographical reference point. Due to the lack of enough genetic information, this population was not considered in the following analyses.

Burunge (~13,000). Another very poorly sampled population, speaking an Afro-Asiatic language (Burunge, [bds]) and living in Tanzania. The town of Kondoa, in the Kondoa district, was chosen as geographical reference point. Due to the lack of enough genetic information, this population was not considered in the following analyses.

Turu (*Turu*; ~550,000). A poorly-sampled population speaking a Bantu language (Nyaturu or Rimi [rim]), living in Tanzania. The Singida town, in the Singida district, was chosen as the geographical reference point. The ALFRED PO000041F ("Bantu speakers") and HDGP "Bantu" populations can be used.

Northeastern Bantu (*Kikuyu*; ~5,300,000). The speakers of a Bantu language (Gikuyu or Kikuyu [kik]), living in Kenya, with the city of Nairobi, the Kenyan capital, as its geographical reference point, was chosen to stand for this ambiguous sample. Genetically, it is represented by the ALFRED PO000058N population, and also ALFRED PO000041F ("Bantu speakers") and HDGP "Bantu" populations can be used.

Biaka Pygmy (*Biaka*; ~5,000-28,000). Pygmy groups, living in Cameroon and Gabon (different from the Baka of DRC and Sudan) and speaking a Bantu language (Yaka or "Babinga", [axk]), but due to information availability, also data from Lingala [lin] was used. Nola, the capital of the Sangha-Mbaéré economic prefecture of the Central African Republic, was chosen as the geographical reference point. Genetically, ALFRED PO000005F and HDGP "Biaka_Pymies" population represent this group.

Zime (*Zime*; ~35,000). Another very poorly sampled population, speaking an Afro-Asiatic language (Pévé, Lamé or Zime, [lme]) and living in Chad and Cameroon. The town of Garoua, capital of the Northern Province of Cameroon, was chosen as geographical reference point. Due to the lack of enough genetic information, this population was not

considered in the following analyses.

Bakola Pygmy (*Bakola*; ~4,000). Another very poorly sampled population, speaking a Bantu language (Gyele or Bakola [gyi]) and living in Cameroon and Equatorial Guinea. The sea port of Kribi in Cameroon, was chosen as geographical reference point. ALFRED PO000041F ("Bantu speakers") and HDGP "Bantu" populations can be used.

Bamoun (*Bamoun*; ~215,000). Another very poorly sampled population, speaking a Bantu language (Bamun or Bamoun [bax]) and living in Cameroon. The town of Foumban in Cameroon, was chosen as geographical reference point. ALFRED PO000041F ("Bantu speakers") and HDGP "Bantu" populations can be used.

Yoruba (*Yoruba*; ~19,000,000). Yoruba are numerous and live in Nigeria and Benin, speaking a Niger-Congo language (Yoruba or Yariba [yor]). The city of Ibadan (Èbá-Ọdàn), the capital of the Ọyọ state of Nigeria and the largest city in Africa, was chosen as the geographical reference point. ALFRED PO000036J and HDGP "Yoruba" populations correspond to this sample.

Mandenka (*Mandenka*; ~1,200,000). A population speaking a Niger-Congo language (Mandinka or Mande [mnk] – but due to lack of information, supplemented with data from Bamanankan or Bambara [bam]), living in Senegal, Gambia and Guinea-Bissau. The city of Ziguinchor, capital of the Casamance region of Senegal, was chosen as the geographical reference point. ALFRED PO000543M and HDGP "Mandenka" populations correspond to this sample.

North Africa and the Near East: 4 populations belong to this geographical region, well studied, both genetically and linguistically.

Mozabite (*Mozabite*, مزاب, M'zab; ~70,000). Small but culturally vigorous population living in the Mzab region of Algeria, dispersed around 7 oases, and speaking an Afro-Asiatic language (Tumzabt or Mzab, [mzb]). The town of Ghardaia, capital of the province with the

same name in Algeria, was chosen as geographical reference point. Genetically, ALFRED PO000570M and HDGP "Mozabite" populations represent this sample.

Druze (*Druze*, درزي, Druse; ~450,000-2,300,000). The Druze are a community living in Israel, Syria, Lebanon and Jordan, characterized by specific religious beliefs, being the descendants of the Isma`ilis (الإسماعيليون), and speaking an Arabic (عربي) dialect [apc]. The city of Haifa (חֵיפָה, حَيْفَا) in northern Israel was chosen as the geographical reference point. Genetically, ALFRED PO000008I and HDGP "Druze" populations represent this sample.

Palestinian (*Palestinian*, فلسطين, Palestinians; ~10,000,000). The Palestinians are a community living in Israel, Egypt, Syria and Jordan, and speaking an Arabic (عربي) dialect [ajp]. The city of Jerusalem (יְרוּשָׁלַיִם, القُدس), capital of Israel, was chosen as the geographical reference point, due to geographical positioning only. Genetically, ALFRED PO000572O and HDGP "Palestinian" populations represent this sample.

Bedouin (*Bedouin*, بدوي, badawī; ~170,000 in Israel). The Bedouin is a generic name applied to desert-living Arab nomads, living in a geographical band extending from the Atlantic coast of North Africa (Sahara) to the eastern coast of the Arabian desert, and speaking an Arabic (عربي) dialect [ayl]. The city of Rahat (רָהָט, رهط) in the south district of Israel, was chosen as the geographical reference point for the Bedouins living in Israel. Genetically, ALFRED PO000571N and HDGP "Bedouin" populations represent this sample.

Asia (Pakistan): 8 populations belong to this geo-political region, well studied, both genetically and linguistically, and quite diverse.

Hazara (*Hazara*; ~9,000,000). Most represented in Afghanistan, but also in Pakistan and Iran, speaking an Indo-European language (Hàzaragi, Azargi or Hazara [haz]). The city of Quetta (كويتہ), capital of the Balochistan province of Pakistan, is a highly multicultural city in a multicultural region, and was chosen as geographical reference point for 3 populations,

including Hazara. Genetically, ALFRED PO000575R and HDGP "Hazara" populations represent this sample.

Balochi (*Balochi*, بلوچ Baloch, Balush; ~5,000,000-6,000,000). Inhabitants of the Balochistan region spanning Iran, Pakistan and Afghanistan, and also in India, speak an Indo-European language (Balochi or Baluchi [bgp]). The city of Quetta (کویتہ), capital of the Balochistan province of Pakistan, is a highly multicultural city in a multicultural region, and was chosen as geographical reference point for 3 populations, including Balochi. Genetically, ALFRED PO000574Q and HDGP "Balochi" populations represent this sample.

Pathan (*Pathan*, پښتون , پختون, Pushtun; ~40,000,000-45,000,000). A group living in Afghanistan and Pakistan, speakers of an Indo-European language (Pashto or Mahsudi [pst]). The city of Quetta (کویتہ), capital of the Balochistan province of Pakistan, is a highly multicultural city in a multicultural region, and was chosen as geographical reference point for 3 populations, including Pathan. Genetically, ALFRED PO000355N and HDGP "Pathan" populations represent this sample.

Burusho (*Burusho*; ~87,000). A small population living in Pakistan and speaking a linguistic isolate (Burushaski, Burushaki, Biltum or Khajuna [bsk]). The generic region of Baluchistan in Pakistan was chosen as the geographical reference. Genetically, ALFRED PO000450J and HDGP "Burusho" populations represent this sample.

Makrani (*Makrani*; ~3,400,000). Population living mainly in Pakistan and Iran, but also in Oman and Arab United Emirates, speaking an Indo-European language (a dialect of Balochi, Makrani [bcc]). The town of Gwadar in the Pakistan Baluchistan region and capital of the Gwadar district, was chosen as geographical reference point. Genetically, the HDGP "Makrani" population represents this sample.

Brahui (*Brahui*; ~2,200,000). A population living in the Kalat district of Pakistan and speaking a Dravidian language (Brahui [brh]). The town of Kalat, the capital of the Kalat district, was chosen as geographical reference point. Genetically, ALFRED PO000573P and HDGP "Brahui" populations represent this sample.

Kalash (*Kalash*, Kalasha of Chitral, Kalasha or Kasivo; ~5,000). Small population living in the Hindu Kush region of Pakistan, phenotypically distinct from its neighbors, and speaking an Indo-European language (Kalasha, Kalashamon [kls]). The Balanguru town in the North-West Frontier Province of Pakistan was chosen as geographical reference point. Genetically, ALFRED PO000449R and HDGP "Kalash" populations represent this sample.

Sindhi (*Sindhi*; ~21,000,000). The Sindhi of Pakistan live mainly in the Sindh region and speak an Indo-European language (Sindhi [snd]). The city of Karachi (كراچي), the capital of the Sindh province, was chosen as geographical reference point. Genetically, ALFRED PO000576S and HDGP "Sindhi" populations represent this sample.

Asia: 18 populations belong to this very diverse geographical region, both genetically and linguistically. Most of these populations (15) live in China.

Hezhen (*Hezhen*, Nanai, нани "Nani", нанайцы "Nanaitsy", 赫哲族 "Hèzhézú"; ~5,700). A population living along the Amur, Sunggari and Ussuri rivers in Russia and China, speaking an Altaic language (Nanai, Gold, Sushen, Hezhen or Hezhe [gld]). The city of Harbin (哈爾濱), capital of the Heilongjiang Province in north-east China, was chosen as geographical reference point. Genetically, ALFRED PO000579V and HDGP "Hezhen" populations represent this sample.

Mongola (*Mongola*, Mongols, Монгол; ~3,300,000). Large population inhabiting mainly Mongolia, China and Russia and speaking an Altaic language (Mongolian, Mongol or Menggu [mvf]). The city of Hohhot (呼和浩特, Хөх хот, Hūhéhàotè), capital of the Inner Mongolian Autonomous Region of China, was chosen as geographical reference point. Genetically, ALFRED PO000502H and HDGP "Mongola" populations represent this sample.

Daur (*Daur*, 达斡 族Dáwò'ěrzú; ~95,000). Small population living in the Inner Mongolia, Heilongjiang and Xinjiang regions of China and Mongolia and speaking an Altaic language (Daur, Dagur, Dawar or Tahur [dta]). The town of Nirji, capital of the Morin Dawa Daur

Autonomous Banner (County) of China, was chosen as geographical reference point. Genetically, ALFRED PO000578U and HDGP "Daur" populations represent this sample.

Orogen (*Orogen* – probably a spelling mistake in the original papers [Evans *et al.* (2005) and Mekel-Bobrov *et al.* (2005)] – Oroqen, 鄂 春族, èlúnchūn zú, Oroqin, Orochen; ~1,200). A very small population in China, speaking an Altaic language (Oroqen, Orochon, Elunchun [orh]). The town of Alihe in China, was chosen as geographical reference point. Genetically, ALFRED PO000541K and HDGP "Oroqen" populations represent this sample.

Miaozu (*Miaozu*, Hmong, 苗, Miáo, Mèo, H'Môngแม้ว "Maew" or ม้ง "Mong"; ~20,000). Population living mainly in China, Vietnam and Laos, and speaking a Hmong-Mien language (Hmong, Guiyang Miao, Miao [hmy]). The province of Guizhou (□ 州, Gùizhōu, Kweichow) in southern China, was chosen as geographical reference point. Genetically, ALFRED PO000487T and HDGP "Miaozu" populations represent this sample.

Yizu (*Yizu*, Yi, Nuosu, 彝族, Yìzú, Lô Lô; ~35,000). Large population living mainly in China and Vietnam, speaking a Sino-Tibetan language (Yi, Ache [yif]). The Minjian town in the Mabian Yi Autonomous County, Sichuan, China, was chosen as geographical reference point. Genetically, ALFRED PO000577T and HDGP "Yizu" populations represent this sample.

Tujia (*Tujia*, Bizika, 土家族; ~70,000). Population inhabiting the Hunan and Hubei provinces of China and speaking a Sino-Tibetan language (Tujia, Tuchia [tji]). The city of Jishou (吉首, Jíshǒu) in the Xiangxi Tujia and Miao Autonomous Prefecture in Hunan province, China, was chosen as geographical reference point. Genetically, ALFRED PO000486S and HDGP "Tujia" populations represent this sample.

Han (*Han*, Han Chinese, 漢族, hànzú; ~1,000,000,000). The largest ethnic group in the world, living mainly in China, it highlights better than anyone else the inherent difficulties of sampling. They speak a Sino-Tibetan language (Chinese, Mandarin Chinese, Mandarin, Guanhua, Beifang Fangyan, Guoyu, Standard Chinese, Putonghua, Hanyu [cmn]). The city of Beijing (北京, Běijīng), capital of China, was chosen as geographical reference point

(highlighting the inappropriateness of using a single point for such a large-scale ethnic structure). Genetically, ALFRED PO0000009J and HDGP "Han" populations represent this sample.

Xibo (*Xibo*, Xibe, 錫伯; ~30,000). Small population living mainly in northeastern China and speaking an Altaic language (Xibe, Sibo, Sibin [sjo]). The city of Shenyang (瀋陽, Shěnyáng), capital of the Liaoning province of China, was chosen as geographical reference point. Genetically, ALFRED PO000580N and HDGP "Xibo" populations represent this sample.

Uygur (*Uygur*, Uighur, ئۇيغۇر, 維吾爾, Wéiwú'ěr, Uyghur; ~7,000,000). A population living in China, Kazakhstan, Kyrgyzstan, Uzbekistan, Turkey, Russia and speaking an Altaic language (Uyghur, Uighur, Wiga [uig]). The city of Urumqi (Ürümqi, ئۈرۈمچى, 烏魯木齊, Wūlǔmùqí), capital of the Xinjiang Autonomous Region of China, was chosen as geographical reference point for the Uygur population of China. Genetically, ALFRED PO000399V and HDGP "Uygur" populations represent this sample.

Dai (*Dai*, Thai, Tai; ~350,000). Population living in southern Yunnan province of China as well as in Laos, Vietnam, Thailand and Myanmar and speaking a Tai-Kadai language (Tai Nüa, Dai Nuea, Tai-Le, Tai-Kong [tdd]). The city of Jinghong (景洪, Jǐnghóng, เชียงรุ้ง), capital of the Xishuangbanna Dai Autonomous Prefecture, Yunnan province of China, was chosen as geographical reference point for the Dai population of China. Genetically, ALFRED PO000464O and HDGP "Dai" populations represent this sample.

Lahu (*Lahu*, 拉祜族, Lāhùzú, Ladhulsi, Kawzhawd, La Hù; ~400,000). Population living in south-east Asia, including China (Yunnan province), Vietnam, Burma, Lahu and Thailand, and speaking a Sino-Tibetan language (Lahu, Lohei, Laku, Kaixien, Namen, Mussuh [lhu]). The city of Kunming (昆明, Kūnmíng), capital of the Yunnan province of China, was chosen as geographical reference point for the Lahu population of China. Genetically, ALFRED PO0000581O and HDGP "Lahu" populations represent this sample.

She (*She*, □ ; ~911). Small population living in China (especially the Fujian, Zhejiang,

Jiangxi, Guangdong and Anhui provinces), and speaking a Hmong-Mien language (She, Huo Nte [shx]). The city of Fuzhou (福州, Fúzhōu, Foochow, Fuchow or Rongcheng, 榕城), capital of the Fujian province of China, was chosen as geographical reference point. Genetically, ALFRED PO000582P and HDGP "She" populations represent this sample.

Naxi (*Naxi*, Nakhi, □ 西族, Nàxī Zú; ~300,000). Small population living in China (Yunnan and Sichuan provinces), and speaking a Sino-Tibetan language (Naxi, Nahsi, Nasi, Nakhi, Lomi, Mu [nbf]). The city of Lijiang (□ 江市, Lìjiāngshì), an administrative division comprising of urban and rural areas in northwestern Yunnan Province of China, was chosen as geographical reference point. Genetically, ALFRED PO000583Q and HDGP "Naxi" populations represent this sample.

Tu (*Tu*, 土; ~150,000). Small population living in China (Qinghai and Gansu provinces), and speaking an Altaic language (Tu, Mongour [mjg]). The town of Xining (西寧, Xīníng), the capital of the Qinghai Province of China, was chosen as geographical reference point. Genetically, ALFRED PO000584R and HDGP "Tu" populations represent this sample.

Cambodian (*Cambodian*, Khmer; ~12,000,000). The predominant ethnic group of Cambodia, also live in Thailand and Vietnam, and speak an Austro-Asiatic (Khmer, Cambodian [khm]). The town of Phnom Penh, the capital of Cambodia, was chosen as geographical reference point. Genetically, ALFRED PO000022E and HDGP "Cambodians" populations represent this sample.

Japanese (*Japanese*, 日本人, Nihon-jin; ~121,000,000). Large population living mainly in Japan, speaking a linguistic isolate (Japanese [jpn]). The city of Tokyo (東京, Tōkyō), the capital of Japan, was chosen as geographical reference point. Genetically, ALFRED PO000010B and HDGP "Japanese" populations represent this sample.

Yakut (*Yakut*, Sakha; ~360,000). Population living in the Sakha (Yakut) Republic of Russia, speaking a n A ltaic language (Yakut, Sakha [sah]). The city of Yakuts k (Яку́тск, Дьокуускай), the capital of the Sakha (Yakut) Republic of Russia, was chosen as geographical reference point. Genetically, ALFRED PO000011C and HDGP "Yakut"

populations represent this sample.

Oceania: 2 populations belong to this geographical region, extremely diverse and interesting from both genetic and linguistic points of view. Unfortunately, due to such a high diversity, coupled with difficulties in sampling, especially on the island of New Guinea, the available samples tend to be be useless for our type of study.

Papuan (~5,600,000 entire Papua-New Guinea). Extremely ambiguous sample, given the lack of specific information (for example, the ALFRED population PO000585S, classified as "Papuan", is described as: "This sample consists of healthy unrelated adult Papuans (Eastern Highlanders) from New Guinea."). Considering the enormous linguistic diversity of New Guinea, two solutions are possible:

1. ignore the Papuan samples;
2. consider that, genetically, the Papuan sample is representative of a highly homogeneous population, so that the entire linguistic diversity must be compared to this assumed genetic uniformity.

Another complication is represented by the possibility that the sample actually comes from cosmopolitan localities and contains a non-negligible proportion of admixed individuals. Therefore, in this study, the first solution was chosen (ignoring the Papuan sample), but only after it was checked that it does not (potentially) impact too much on the conclusions (see main text for.details).

NAN Melanesian (*NANMelanesian*; ~10,000). This population is very poorly specified in the original papers [Evans *et al.* (2005) and Mekel-Bobrov *et al.* (2005)], but it turns out that NAN Melanesian stands for *Non-Austronesian Melanesian*, and the most probable candidate is represented by the *Naasioi* of Bougainville, Papua New Guinea, speaking an East Papuan language (Naasioi, Nasioi, Kieta, Kieta Talk or Aunge [nas]). The island of Bougainville (also known as Papala), the largest of the Solomon islands, was chosen as geographical reference point. Genetically, ALFRED PO000012D and HDGP "Melanesian" populations (with a high probability) represent this sample.

Europe: 8 populations belong to this geographical region, very well sampled and known,

both genetically and linguistically.

French Basque (*FrBasque*; ~250,000 in France). Population living in the Northern Basque Country (French Basque Country, Continental Basque Country, Pays Basque or Iparralde), speaking a linguistic isolate (Basque, Vascuense, Euskera [eus]). The city of Bayonne (Baiona), the main town of Labourd in the French Basque Country, France, was chosen as geographical reference point for the French Basque population. Genetically, ALFRED PO000042G and HDGP "Basques" populations represent this sample.

French (*French*; ~51,000,000 in France). Large population mainly inhabiting France, speaking an Indo-European language (French, Français [fra]). The city of Pris, the capital of France, was chosen as geographical reference point for the French population. Genetically, ALFRED PO000111D and HDGP "French" populations represent this sample.

Sardinian (*Sardinian*; ~1,600,000). Population living on the island of Sardinia (Sardegna, Sardigna or Sardinna), speaking an Indo-European language (Sardinian, Sard, Sardarese, Logudorese [src]). The city of Cagliari (Càgliari, Casteddu), the capital of the Sardinia autonomous region of Italy, was chosen as geographical reference point. Genetically, ALFRED PO000411G and HDGP "Sardinian" populations represent this sample.

North Italian (*NItalian*; ~10,000,000). This population is not very well defined, but was taken to inhabit the north of Italy, speaking an Indo-European language (Venetian, Veneto, Venet [vec]). The city of Bergamo (Bèrghem) was chosen as geographical reference point. Genetically, the HDGP "Bergamo" population represents this sample.

Tuscan (*Tuscan*; ~3,000,000). Population living in the region of Tuscany (Toscana), central Italy, and speaking an Indo-European language (Italian, Italiano [ita]). The city of Florence (Firenze), the capital of Tuscany, Italy, was chosen as geographical reference point. Genetically, ALFRED PO000137L and HDGP "Tuscan" populations represent this sample.

Orcadian (*Orcadian*; ~20,000). Population inhabiting the Orkney Islands, and speaking an Indo-European language (Scots [sco]). The city of Kirkwall, the capital of the Orkney

Islands, Scotland, UK, was chosen as geographical reference point. Genetically, ALFRED PO000586T and HDGP "Orcadians" populations represent this sample.

Russian (*Russian*, Russians, Русские; ~140,000,000). Population living in Russia and neighboring countries (Ukraine, Kazakhstan, Belarus, etc.), and speaking an Indo-European language (Russian, Russki [rus]). The city of Moscow (Москва́, Moskva), the capital of Russia, was chosen as geographical reference point. Genetically, ALFRED PO000019K and HDGP "Russians" populations represent this sample.

Adygei (*Adygei*, Adyghe, Adygs, Cherkess; ~500,000). Population inhabiting the north Caucasus region, mainly the Republic of Adygea (Респу́блика Адыге́я, Адыгэ Республик), but also Republic of Karachay-Cherkessia (Карача́ево-Черке́сская респу́блика), where they are known as Cherkess, both in the Russian Federation, and speaking a North-Caucasian language (Adyghe, Circassian, Kiakh, Kjax, Adygei, Adygey [ady]). The city of Maykop (Майко́п), the capital of the Republic of Adygea, was chosen as geographical reference point. Genetically, ALFRED PO000017I and HDGP "Adygei" populations represent this sample.

Annex 6: Description of the linguistic data

In this annex, the linguistic data (the 28 linguistic features and their values) is described. The gathering of linguistic features values for the various languages involved represented a very tedious process, as there are very few complete works centralizing and systematizing such data (Haspelmath, Dryer, Gil & Comrie (2005) represents an impressive attempt, but, unfortunately, still far from complete). Moreover, given the nature of the data and the diversity (and, sometimes, incongruence) of sources, a certain degree of subjectivity is involved, but we are confident that a team of independent linguists will arrive at results consistent with ours[340].

Annex 6.1: Description of data sources and methods

General observations and data sources: This subsection concerns the languages (populations) and not any particular linguistic feature in special. The main data sources used are Haspelmath, Dryer, Gil & Comrie (2005) and Campbell (2000), to which specific sources for specific languages are added (listed in Table 46). Due to information availability issues, the following *replacement conventions* were applied:

- for the Arabic dialects (apc, ajp and ayl) information from the better covered Spoken Egyptian Arabic (arz) was used;
- for Makrani (bcc), information from its better documented dialect Balochi (bgp) was used;
- for Mongola (mvf), information from its better known dialect (khk) was used;
- for Orogen/Oroqen (orh), information from the related Evenki (evn) was used;
- for Miaozu (hmy), information from the related Hmong Njua (blu) was used;
- for Xibo (sjo), information from the related Manchu (mnc) was used;
- for Dai (tdd), information from the related Thai (tha) was used;
- for Orcadian (sco), information from related English (eng) was used;
- for Mbuti Pygmy (les), information from the related Efe [efe] and Ma'di [mhi, smn]

340Where possible, this type of subjective judgments was done by specialists (see details below). A statistical test of inter-evaluator consistency (inter-rater reliability) (e.g., Loewenthal, 2001:14) is certainly feasible and interesting in itself, but given the daunting amount of work involved, it is highly improbable that it will be practically feasible.

was used.

Besides these replacement conventions, *completion conventions* for missing data in the primary language were also applied:

- for Turu (rim), data completed from related Langi/Rangi (lag);
- for Biaka Pygmy (axk), data completed from related Lingala (lin);
- for Mandenka (mnk), data completed from related Bamanankan (bam);
- for Hazara (haz), data completed from related Persian/Farsi (pes).

Most personal communications (*pc*) are based on a standardized questionnaire realized by prof. D.R. Ladd and sent by him through e-mail to linguists specialized in relevant areas. If "D. R. Ladd" is not specified for a source, then the data collection from that source is due to me (Dan Dediu).

Population (language)	***Sources***
SESWBantu	Haspelmath, Dryer, Gil & Comrie (2005) D. R. Ladd (*pc*)
San	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Mbuti	D. R. Ladd (Nigel Fabb, Mairi Blackings, *pc*)
Masai	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (Tucker & Mpaayei 1955)
Sandawe	Haspelmath, Dryer, Gil & Comrie (2005) D. R. Ladd (Helen Eaton, *pc*)
Burunge	Haspelmath, Dryer, Gil & Comrie (2005) D. R. Ladd (Oliver Stegen, Michael Endl, *pc*)
Turu	D. R. Ladd (Oliver Stegen, *pc*)
Kikuyu	Haspelmath, Dryer, Gil & Comrie (2005) D. R. Ladd (Mugane, 1997)
Biaka	D. R. Ladd (Guthrie, 1948; Guthrie, 1953)
Zime	D. R. Ladd (Jim Roberts, *pc*)
Bakola	D. R. Ladd (Koen Bostoen, *pc*; Guthrie, 1948; Guthrie, 1953)
Bamoun	D. R. Ladd (Bruce Connell, *pc*; Guthrie, 1948; Guthrie, 1953)
Yoruba	Haspelmath, Dryer, Gil & Comrie (2005)

Population (language)	*Sources*
	D. R. Ladd (*pc*)
Mandenka	Haspelmath, Dryer, Gil & Comrie (2005) D. R. Ladd (*pc*)
Mozabite	D. R. Ladd (Rachid Ridouane, *pc*; Penchoen, 1973)
Druze **Palestinian** **Bedouin**	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) Dan Dediu (Jim Hurford, *pc*)
Hazara	D. R. Ladd (Lazard, 1992)
Balochi	D. R. Ladd (Schmitt (ed.), 1989: esp. Josef Elfenbein)
Pathan	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (Schmitt (ed.), 1989: esp. Prods O. Skjaervø)
Burusho	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Makrani	D. R. Ladd (see Balochi)
Brahui	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (Bashir, 1991)
Kalash	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (Masica, 1991)
Sindhi	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (Masica, 1991)
Hezhen	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Mongola	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Daur	Haspelmath, Dryer, Gil & Comrie (2005)
Orogen	Xi (1996:136) Haspelmath, Dryer, Gil & Comrie (2005)
Miaozu	Mortensen (2004) Haspelmath, Dryer, Gil & Comrie (2005)
Yizu	Campbell (2000) D. R. Ladd (Jerry Edmondson, Lama Ziwo, *pc*)
Tujia	D. R. Ladd (Jerry Edmondson, Lama Ziwo, *pc*)

Population (language)	*Sources*
Han	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (*pc*)
Xibo	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Uygur	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (*pc*)
Dai	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Lahu	Haspelmath, Dryer, Gil & Comrie (2005)
She	D. R. Ladd (Jerry Edmondson, Lama Ziwo, *pc*)
Naxi	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000) D. R. Ladd (Jerry Edmondson, Lama Ziwo, *pc*)
Tu	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Cambodian	Haspelmath, Dryer, Gil & Comrie (2005)
Japanese	Haspelmath, Dryer, Gil & Comrie (2005)
Yakut	Haspelmath, Dryer, Gil & Comrie (2005)
Papuan	Haspelmath, Dryer, Gil & Comrie (2005) In this case, judgments are based on the most frequent value given by Haspelmath, Dryer, Gil & Comrie (2005) for the entire eastern half of the island.
NANMelanesian	Organised Phonology Data: Nasioi [government spelling] (Naasioi [language spelling]) Language [NAS] Kieta – North Solomons Province Haspelmath, Dryer, Gil & Comrie (2005)
FrBasque	Haspelmath, Dryer, Gil & Comrie (2005)
French	Haspelmath, Dryer, Gil & Comrie (2005)
Sardinian	D. R. Ladd (*pc*)
NItalian	D. R. Ladd (*pc*)
Tuscan	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)
Orcadian	Haspelmath, Dryer, Gil & Comrie (2005)
Russian	Haspelmath, Dryer, Gil & Comrie (2005)

Population (language)	*Sources*
Adygei	Haspelmath, Dryer, Gil & Comrie (2005) Campbell (2000)

Table 46: The sources used for gathering the linguistic features per population/language.

The following people are gratefully thanked for their invaluable contribution through personal communications (mostly managed by Prof. D. R. Ladd):

Personal communication source	*Identification information*
Bruce Connell	Dr Bruce Connell, (August 2006) University of Kent, Kent, UK
Constance Kutsch-Lojenga	Dr. Constance Kutsch Lojenga Lecturer/Researcher, Department of Languages and Cultures of Africa, Leiden University, The Netherlands http://www.sil.org/sil/roster/lojenga_constance.htm
D. R. Ladd	Prof. D. R. Ladd Professor of Linguistics, Linguistics and English Language, The University of Edinburgh, UK http://www.ling.ed.ac.uk/~bob/
Helen Eaton	Dr. Helen Eaton SIL International (Tanzania) http://www.drhelenipresume.com/
Jerry Edmondson	Jerold A. Edmondson Professor of Linguistics, University of Texas at Arlington, Arlington, TX, USA http://ling.uta.edu/~jerry/
Jim Hurford	Prof. Jim Hurford Professor of General Linguistics, Linguistics and English Language,

Personal communication source	*Identification information*
	The University of Edinburgh, UK http://www.ling.ed.ac.uk/~jim/
Jim Roberts	Jim Roberts, SIL Chad
Koen Bostoen	Dr. Koen Bostoen Faculté de Philosophie et Lettres, Campus du Solbosch, Bruxelles, Belgium
Lama Ziwo	Dr. Lama Ziwo University of Texas at Arlington, Arlington, TX, USA
Mairi Blackings	Dr. Mairi Blackings University of Strathclyde, UK
Michael Endl	Michael Endl, SIL Tanzania
Nigel Fabb	Prof. Nigel Fabb Professor of Literary Linguistics, University of Strathclyde, UK http://www.strath.ac.uk/english/staff/fabbnigelprof/
Oliver Stegen	Oliver Stegen SIL International (Tanzania) http://www.ling.ed.ac.uk/~oliver/
Rachid Ridouane	Dr. Rachid Ridouane, Ecole Doctorale "Langage et Langues", l'Université Paris 3, Paris, France http://www.univ-paris3.fr/recherche/sites/edll/student/strr/index.html
Ron Asher	Prof. Ronald Asher Emeritus Professor, Linguistics and English Language, University of Edinburgh,

Personal communication source	*Identification information*
	UK

Table 47: The identification information for the personal communications sources.

The 28 linguistic features are given in Table 48. The coding is based on the schemes used by Haspelmath, Dryer, Gil & Comrie (2005)[341] and details are given in the "comments" column[342].

Name	*Description*	*Coding*	*Comments*
ConsCat	The richness of consonant inventory.	0 = small, moderately small & average 1 = moderately large or large	Phonology: Consonant Inventories by Ian Maddieson
Cons*	The actual number of consonants.		D. R. Ladd
VowelsCat	The richness of vowel inventory.	0 = small & average 1 = moderately large or large	Phonology: Vowel Quality Inventories by Ian Maddieson
Vowels*	The actual number of vowels.		D. R. Ladd
UvularC	The presence or absence of uvular consonants	0 = none 1 = uvular stops, uvular continuants or both	Phonology: Uvular Consonants by Ian Maddieson
GlotC	The presence or absence of glottalized consonants	0 = no glottalized consonants 1 = any category of glottalized consonants	Phonology: Glottalized Consonants by Ian Maddieson
VelarNasal	The presence or absence of velar nasals	0 = no velar nasal 1 = initial velar nasal or not initial velar nasal	Phonology: The Velar Nasal by Gregory D. Anderson
FrontRdV	The presence or absence of front rounded vowels	0 = none 1 = high, mid or both	Phonology: Front Rounded Vowels by Ian Maddieson

341 The accompanying interactive reference tool (WALS Software): therefore, all reference in the "comments" column are given relative to this software.

342 Some features are inherently impossible to encode binary in a meaningful way, leading to the decision to mark certain cases as missing data (e.g., the word order features for languages without a dominant word order). This decision leads to a minimal biasing of the data.

Name	*Description*	*Coding*	*Comments*
Codas*	Are codas allowed?	0 = no codas allowed 1 = otherwise	D. R. Ladd
OnsetClust*	Are onset cluasters allowed?	0 = no onset clusters allowed 1 = otherwise	D. R. Ladd
WALSSylStr	The complexity of syllable structure.	0 = simple or moderatetly complex 1 = complex	Phonology: Syllable Structure by Ian Maddieson
Tone	Does the language have a tonal system?	0 = no tones 1 = simple or complex tonal systems	Phonology: Tone by Ian Maddieson
RareC	Does the language have any rare consonants?	0 = none 1 = clicks, labial-velar, pharyngeals or 'th' sounds	Phonology: Presence of Uncommon Consonants by Ian Maddieson
Affixation	How much affixation does the language use?	0 = little affixation 1 = strongly suffixing, weakly suffixing, equal suffixing and prefixing, weakly prefixing or strong prefixing	Morphology: Prefixing vs. Suffixing in Inflectional Morphology by Matthew S. Dryer
CaseAffixes	Are cases marked with affixes?	0 = yes 1 = no case affixes or adpositional clitics	Nominal Categories: Position of Case Affixes by Matthew S. Dryer
NumClassifiers	Does the language have numeral classifiers?	0 = no 1 = optional or obligatory	Nominal Categories: Numeral Classifiers by David Gil
TenseAspect	Are there tense-aspect marking inflections?	0 = no tense-aspect inflection 1 = tense-aspect prefixes, tense-aspect suffixes, tense-aspect tone or mixed type	Verbal Categories: Position of Tense-Aspect Affixes by Matthew S. Dryer
MorphImpv	Does the language have second person imperatives as dedicated morphological categories?	0 = no second person imperatives 1 = second singluar and second plural, second singular, second plural or second person	Verbal Categories: The Morphological Imperative by Johan van der Auwera, Ludo Lejeune (Umarani

Name	*Description*	*Coding*	*Comments*
		number-neutral	Pappuswamy, Valentin Goussev)
SVWO	What is the dominant Subject-Verb word order (if any)?	0 = SV 1 = VS The only languages without dominant SV word order are (src, vec & ita) and were marked as missing data.	Word Order: Order of Subject and Verb by Matthew S. Dryer
OVWO	What is the dominant Object-Verb word order (if any)?	0 = OV 1 = VO The only language without dominant SV word order is (efe/mhi) and was marked as missing data.	Word Order: Order of Object and Verb by Matthew S. Dryer
AdposNP	What is the dominant order (if any) between adposition and noun phrase?	0 = postpositions 1 = prepositions For Indo-Aryan languages (bgp, pst, bsk, bcc, brh, kls & snd), it seems there is a certain degree of dominance of 0, even if the situation is complex (e.g., Campbel, 2000). For Burunge (bds), prepositions are prefered.	Word Order: Order of Adposition and Noun Phrase by Matthew S. Dryer
GenNoun	What is the dominant order (if any) between genitive and noun?	0 = genitive-noun 1 = noun-genitive The languages without dominant order (lhu, efe/mhi & nas) were marked as missing data.	Order: Order of Genitive and Noun by Matthew S. Dryer.
AdjNoun	What is the dominant order (if any) between adjective and noun?	0 = adjective-noun 1 = noun-adjective The only language without dominant order (nas) was marked as missing data.	Word Order: Order of Adjective and Noun by Matthew S. Dryer
NumNoun	What is the dominant	0 = numeral-noun	Word Order: Order of

Name	*Description*	*Coding*	*Comments*
	order (if any) between numeral and noun?	1 = noun-numeral The only language without dominant order (ady) was marked as missing data. The case of Arabic dialects was marked 0, given that the only exception is the numeral "one" (Jim Hurford, *pc*)	Numeral and Noun by Matthew S. Dryer
InterrPhr	Is the interrogative phrase initial?	0 = not initial interrogative phrase 1 = initial interrogative phrase	Word Order: Position of Interrogative Phrases in Content Questions by Matthew S. Dryer
Passive┼	Is there a passive construction?	0 = absent 1 = present We have a very low confidence in its meaning in various sources.	Simple Clauses: Passive Construction by Anna Siewierska
NomLoc	A language is called a share-language if the encoding strategy for locational predications is (or can be) used for nominal predications, and a split-language if the encoding strategies for the two constructions must be different.	0 = different (split-language) 1 = identical (share-language)	Simple Clauses: Nominal and Locational Predication by Leon Stassen (the definition is also taken from here)
ZeroCopula	Is the omission of copula allowed?	0 = impossible 1 = possible	Simple Clauses: Zero Copula for Predicate Nominals by Leon Stassen

Table 48: The list of the 28 linguistic features with description, coding scheme and comments.

The starred (*) features are original, due to prof. D. R. Ladd. The cross (┼) marks those features in which we (D. R. Ladd and me) don't have a high confidence, due to the ambiguity

Name	*Description*	*Coding*	*Comments*
	order (if any) between numeral and noun?	1 = noun-numeral The only language without dominant order (ady) was marked as missing data. The case of Arabic dialects was marked 0, given that the only exception is the numeral "one" (Jim Hurford, *pc*)	Numeral and Noun by Matthew S. Dryer
InterrPhr	Is the interrogative phrase initial?	0 = not initial interrogative phrase 1 = initial interrogative phrase	Word Order: Position of Interrogative Phrases in Content Questions by Matthew S. Dryer
Passive†	Is there a passive construction?	0 = absent 1 = present We have a very low confidence in its meaning in various sources.	Simple Clauses: Passive Construction by Anna Siewierska
NomLoc	A language is called a share-language if the encoding strategy for locational predications is (or can be) used for nominal predications, and a split-language if the encoding strategies for the two constructions must be different.	0 = different (split-language) 1 = identical (share-language)	Simple Clauses: Nominal and Locational Predication by Leon Stassen (the definition is also taken from here)
ZeroCopula	Is the omission of copula allowed?	0 = impossible 1 = possible	Simple Clauses: Zero Copula for Predicate Nominals by Leon Stassen

Table 48: The list of the 28 linguistic features with description, coding scheme and comments.

The starred (*) features are original, due to prof. D. R. Ladd. The cross (†) marks those features in which we (D. R. Ladd and me) don't have a high confidence, due to the ambiguity

of their meaning and differing interpretations given to them in different sources.

Annex 6.2: The values of the 28 linguistic features for each of the 54 populations of the OWF sample

The following table (Table 49) contains the values of each of the 28 linguistic features for each of the 54 populations (languages) of the *OWF sample*. The features are binary ("0" or "1"), missing data are represented by empty cells. The case of *Tone* for *Papuan* ("0&1") is a real ambiguity, due to the fact that almost equal proportions of languages spoken in Papua-New Guinea highlands have or do not have tones. Therefore, this case needs special treatment during data analysis (see main text for details).

Population	*ConsCat*	*Cons*	*VowelsCat*	*Vowels*	*UvularC*	*GlotC*	*VelarNasal*	*FrontRdV*	*Codas*	*OnsetClust*	*WALSSylStr*	*Tone*	*RareC*	*Affixation*	*CaseAffixes*	*NumClassifiers*	*TenseAspect*	*MorphImpv*	*SVWO*	*OVWO*	*AdposNP*	*GenNoun*	*AdjNoun*	*NumNoun*	*InterrPhr*	*Passive*	*NomLoc*	*ZeroCopula*
SESWBantu	1	35	0	5	0	1	0	0	0	0	0	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	0
San	1	31	0	5	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
Mbuti	1	39	1	9	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0		0		1	1	0	0	1	1
Masai	0	18	1	9	0	1	1	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0
Sandawe	1	44	0	5	0	1	0	0	0	0	0	1	1	1	0	0		1	0	0	0	0	1	1	1	0	0	1
Burunge	1	30	0	5	1	1	0	0	0	0	0	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	0	1
Turu	1	30	1	7	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	1	1	1	1	1	1	1	0	0
Kikuyu	0	18	1	7	0	0	1	0	0	1	0	1	0	1	0		1	1	0	1	1	1	1	1	0	1		
Biaka	0	25	1	7	0	1	1	0	0	0	0	1	1	1	0	0	1	1	0	1	1	1	1	1	0	1		1
Zime	1	25	0	5	0	1	1	0	1	0	0	1	0	1	0	0	1		0	1	1	1	1	1	0	0		1
Bakola	0	17	1	7	0	1	0	0	0	0	0	1	0	1	0	0	1		0	1		1	1	1	0	1		1
Bamoun	0	16	1	8	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	1	1	0	1	1	0	1	0	0
Yoruba	0	17	1	7	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	1
Mandenka	0	19	1	7	0	0	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	1	1	0	0	0	0
Mozabite	0	19	0	3	1	0	0	0	1	1	1	0	1	1	0		1	1	1	1	1	1	1	0	1	1	0	1
Druze	0	27	0	3	1	0	0	0	1	0	1	0	1	1	0	0	1	1	0	1	1	1	1	0	0	1	1	1
Palestinian	0	27	0	3	1	0	0	0	1	0	1	0	1	1	0	0	1	1	0	1	1	1	1	0	0	1	1	1
Bedouin	0	27	0	3	1	0	0	0	1	0	1	0	1	1	0	0	1	1	0	1	1	1	1	0	0	1	1	1
Hazara	0	23	0	6	1	0	0	0	1	0	1	0	0	1	0	0	1	1	0	0	1	1	1	0	0	1	1	0
Balochi	0	24	0	5	1	0	0	0	1	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0		1	1	0
Pathan	1	29	0	7	1	0	1	0	1	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	0

Population	*ConsCat*	*Cons*	*VowelsCat*	*Vowels*	*UvularC*	*GlotC*	*VelarNasal*	*FrontRdV*	*Codas*	*OnsetClust*	*WALSSylStr*	*Tone*	*RareC*	*Affixation*	*CaseAffixes*	*NumClassifiers*	*TenseAspect*	*MorphImpv*	*SVWO*	*OVWO*	*AdposNP*	*GenNoun*	*AdjNoun*	*NumNoun*	*InterrPhr*	*Passive*	*NomLoc*	*ZeroCopula*
Burusho	1	30	0	5	1	0	1	0			1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	0
Makrani	0	25	0	8	1	0	0	0	1	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0		1	1	0
Brahui	0	25	0	10	0	0	0	0	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	1	0
Kalash	1	38	1	8	1	0	0	0	1	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0		1
Sindhi	1	30	1	10	0	1	1	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	0
Hezhen	0	19	0	6	0	0	1	0			0	0	0			0	1	1	0	0	0						1	1
Mongola	0	14	1	7	0	0	1	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1
Daur	0		1		0	0	1	1			1	0	0	1	1		1	1	0	0	0	0	0	0	0	1		
Orogen	0	25	0	10	0	0	1	0			0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	1	1	1
Miaozu	1	57	0	9	1	0	1	0			0	1	0	0	0		0	0	0	1	1	0	1	0	0	1		
Yizu	1	44	1	10	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1
Tujia	0	21	1	7	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1
Han	1	24	0	6	1	0	1	1	1	0	0	1	0	0	0	1	1	0	0	1	1	0	0	0	0	1	0	0
Xibo	1	22	0	6	0	0		1				0	0	1	1		1		0	0	0		0			1	0	0
Uygur	0	25	1	9	1	0	1	1	1	1	1	0	0	1	1		1		0	0	0	0	0	0	1	1		
Dai	0	22	1	9	0	0	1	0			0	1	0	0	0	1	0	0	0	1	1	1	1	1	0	1	0	0
Lahu	1		1		1	0	1	0			0	1	0	0	0	1			0	0	0	0		1	0	0	0	0
She	1	33	1	18	0	0	1	0	1	0	1	1	0	0	0	1	0	0	0	1	1	0	1	0	0	0		0
Naxi	1	32	1	12	1	0	1	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1
Tu	0	22	0	8	1	0	1	0			0	0	0	1	1		1		0	0	0		0	0				
Cambodian	0	20	1	25	0	1	1	0	1	1	1	0	0	0	0	1	0	0	0	1	1	1	1	1	0	0	0	1

Population	*ConsCat*	*Cons*	*VowelsCat*	*Vowels*	*UvularC*	*GlotC*	*VelarNasal*	*FrontRdV*	*Codas*	*OnsetClust*	*WALSSylStr*	*Tone*	*RareC*	*Affixation*	*CaseAffixes*	*NumClassifiers*	*TenseAspect*	*MorphImpv*	*SVWO*	*OVWO*	*AdposNP*	*GenNoun*	*AdjNoun*	*NumNoun*	*InterrPhr*	*Passive*	*NomLoc*	*ZeroCopula*
Japanese	0	13	0	5	1	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0
Yakut	0	20	1	8	0	0	1	1			1	0	0	1	1		1		0	0	0	0	0				1	0
NANMelanesian	0	8	0	5	0	0	1	0	1	0		0	0	1	1	0	1		0	0	0			0	0	0	0	1
FrBasque	0	24	0	5	0	0	0	0	1	0	1	0	0	1	1	0	1	1	0	0	0	0	1	0	0	1	0	0
French	0	19	1	11	1	0	0	1	1	1	1	0	0	1	0	0	1	1	0	1	1	1	1	0	1	1	1	0
Sardinian	0	15	0	5	0	0	0	0	1	1	1	0	0	1	0	0	1	1		1	1	1	1	0	1	1	1	0
NItalian	0	19	0	5	0	0	0		1	1	1	0	0	1	0	0	1	1		1	1	1	1	0	1	1	1	0
Tuscan	0	19	1	7	0	0	1	0	1	1	1	0	0	1	0	0	1	1		1	1	1	1	0	1	1	1	0
Orcadian	0	24	1	11	0	0	1	0	1	1	1	0	1	1	0	0	1	0	0	1	1	0	0	0	1	1	1	0
Russian	1	25	0	5	0	0	0	0	1	1	1	0	0	1	1	0	1	1	0	1	1	1	0	0	1	1	1	1
Adygei	1	80	0	3	1	1	0	0		1	1	0	1	1	1	0	1	1	0	0	0	0	1		1	0	0	0
Papuan	0		0		0	0		0		0	0	0&1	0	1	1	0	1	1	0	0	0	0	1	1	0	0	0	1

Table 49: The values of each of the 28 linguistic features for each of the 54 populations (languages) of the OWF sample.

The Papuan by Tone cell (light gray) reflects a real ambiguity, where bot 0 and 1 values have equal frequencies in the Papua-New Guinean highlands.

Annex 7: Spatial analyses

Annex 7.1: The genetic distance matrices for ASPM and MCPH

The genetic distance matrices for *ASPM* and *MCPH* are represented in the gray-scale encoding in Figures 77 and 78 below.

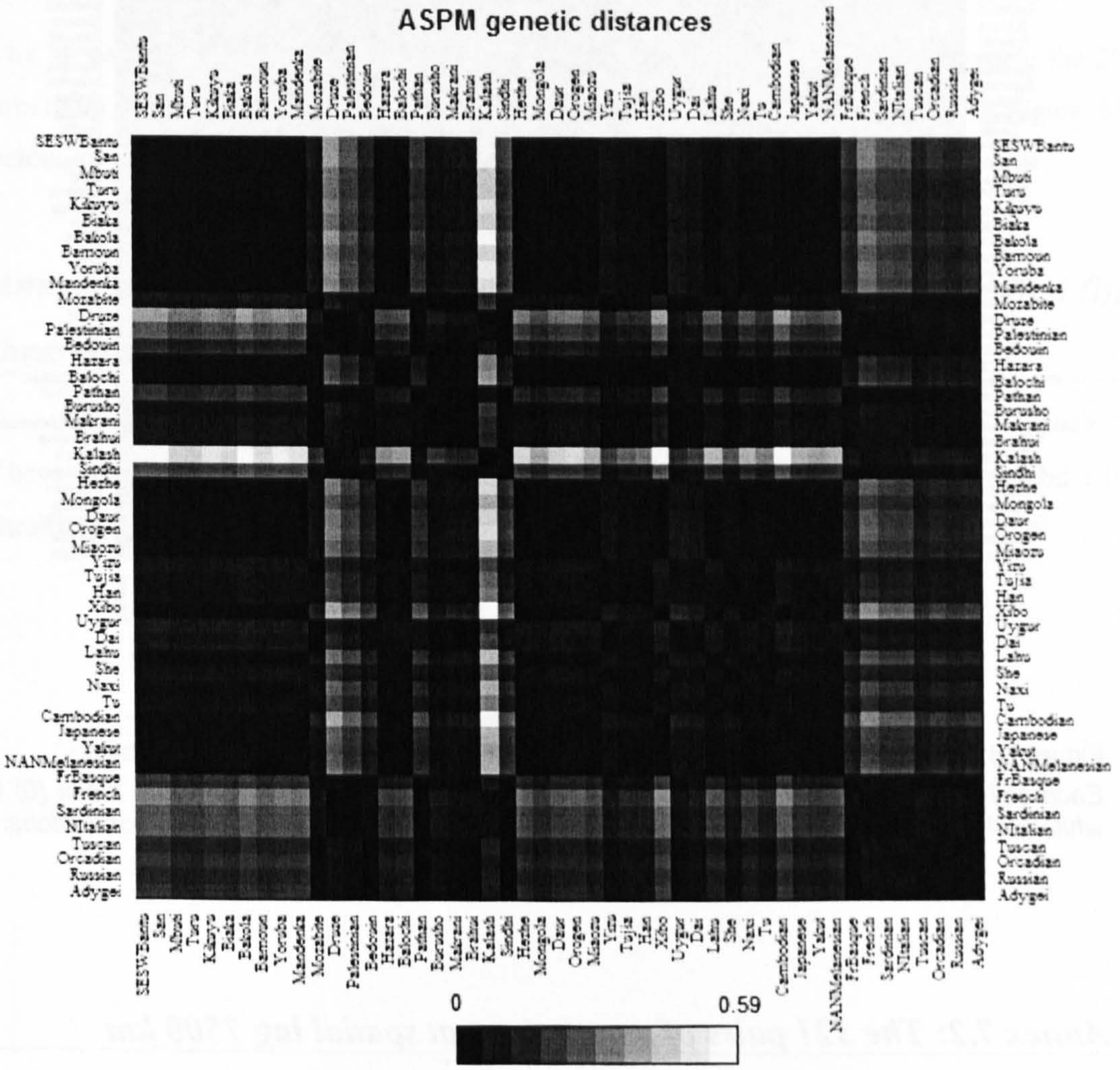

Figure 77: The ASPM genetic distances (Nei's D) matrix in gray-scale representation.

Each cell represents the genetic distance between a pair of populations, from black (0) to white (0.59). The pattern is mostly homogeneous, except for some populations which are differentiated: Kalash, Sindhi, Druze and Palestinian. Striking is the resemblance of European and East/South/Central Asian populations.

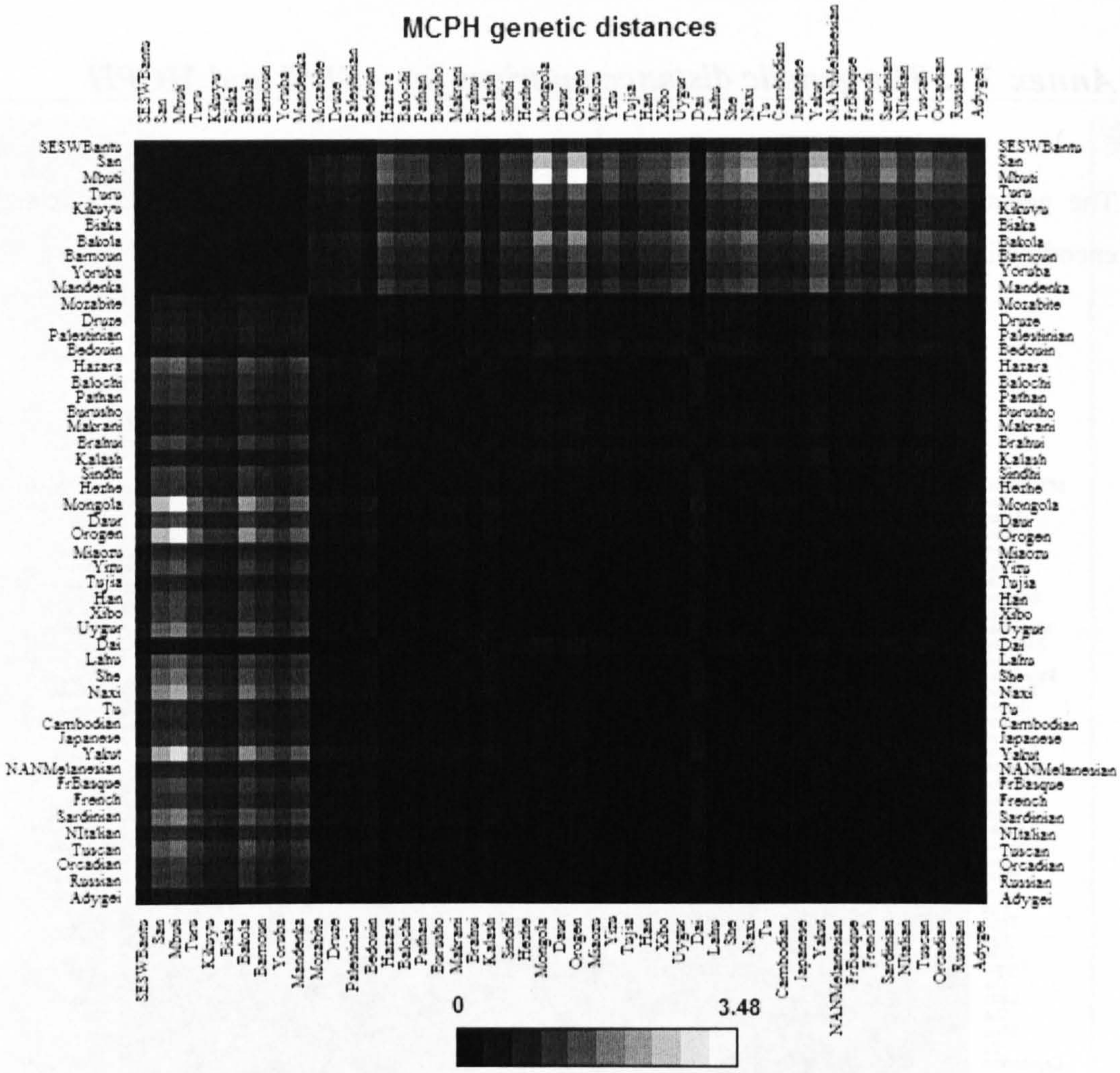

Figure 78: The MCPH genetic distances (Nei's D) matrix in gray-scale representation.
Each cell represents the genetic distance between a pair of populations, from black (0) to white (3.48). The pattern is mostly homogeneous, except for the sub-Saharan populations.

Annex 7.2: The 321 pairs of populations at spatial lag 7500 km

The 321 pairs of populations separated by a spatial lag of 7500 ± 1500 km, where the 1st maximum of *Tone*, *WALSSylStr* and *Codas* occurs (Section 4.7.3), are represented in Figure 79 below.

Annex 7.3: The 65 pairs of populations at spatial lag 13,500 km

The 65 pairs of populations separated by a spatial lag of 13,500 ± 1500 km, where the 2nd minimum of *Tone* occurs (Section 4.7.3), are represented in Figure 80 below.

Annex 7.4: The 30 pairs of populations at spatial lag 15,000 km

The 30 pairs of populations separated by a spatial lag of 15,000 ± 1500 km, where the 2nd minimum of *WALSSylStr* and *Codas* occurs (Section 4.7.3), are represented in Figure 81 below.

Annex 7.5: Geographic, genetic and linguistic boundaries: method (i), thresholds $\tau = .10$ and $\tau = .25$, and method (ii), threshold $\tau = .10$

These boundaries (Section 4.7.4), are represented in Figures 82 – 92. For method (ii), threshold $\tau = .25$, see Section 4.7.4.

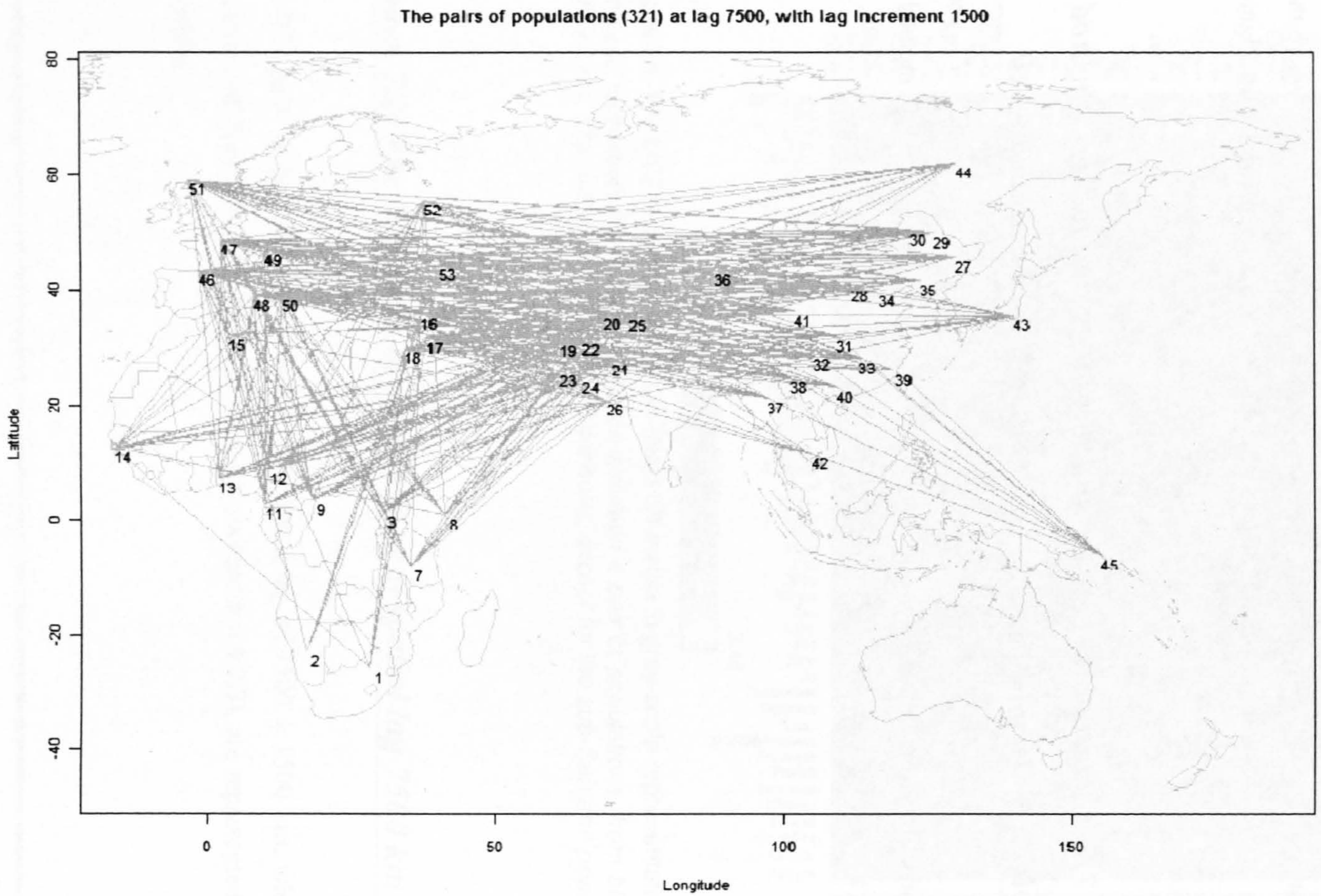

Figure 79: The pairs of populations (321) separated by a lag of 7500 ± 1500km.

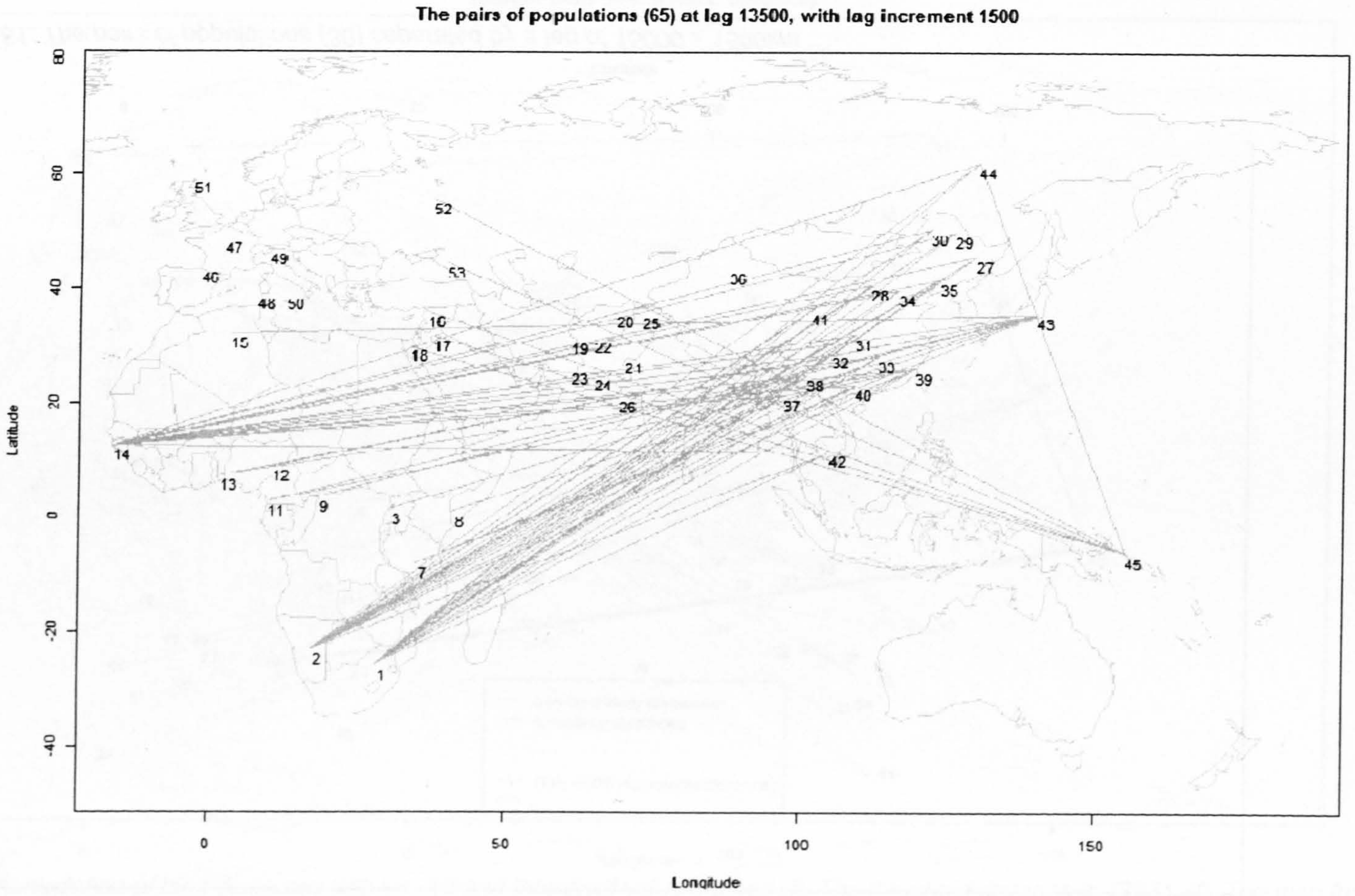

Figure 80: The pairs of populations (65) separated by a lag of 13500 ± 1500km.

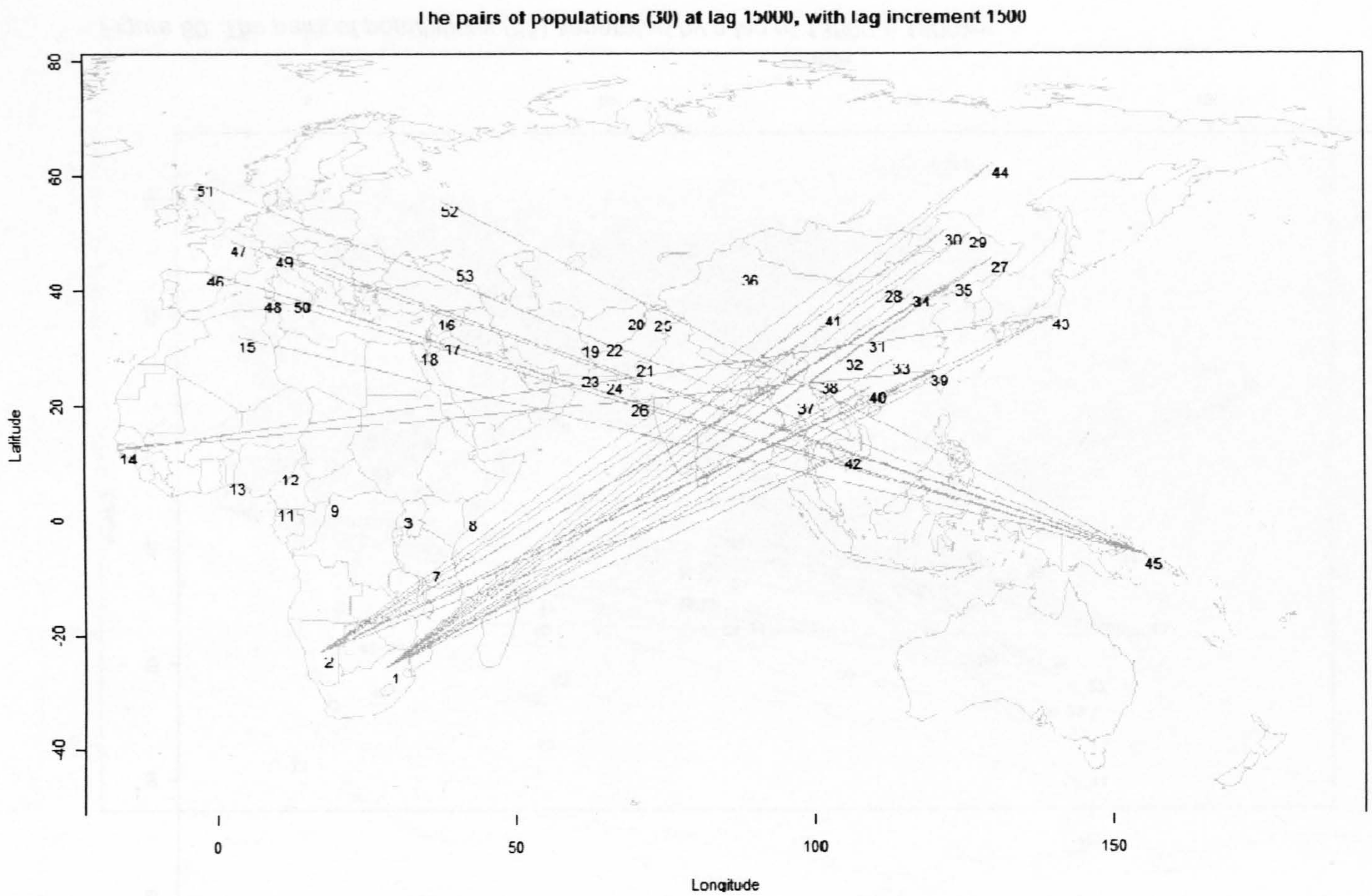

Figure 81: The pairs of populations (30) separated by a lag of 15000 ± 1500km.

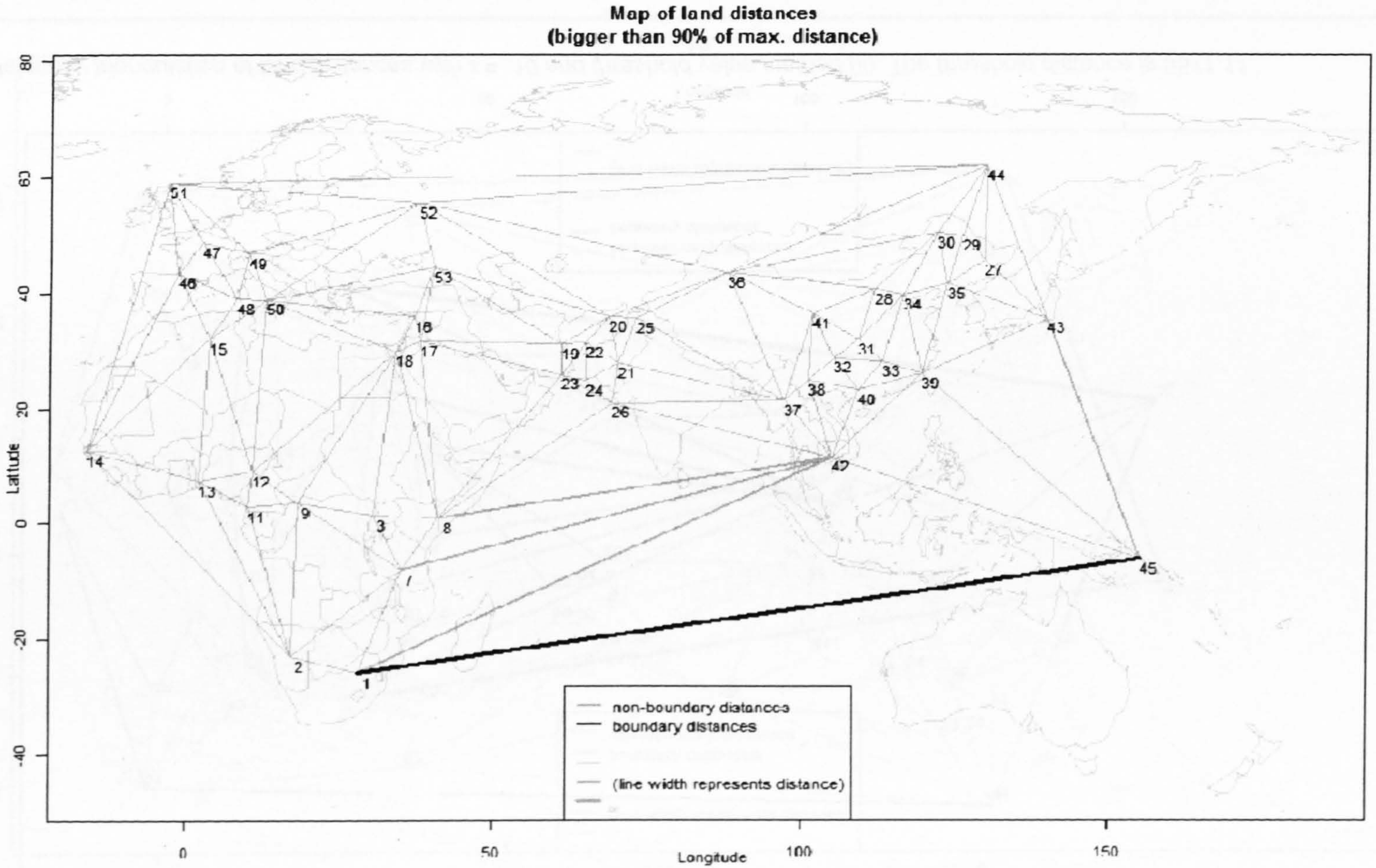

Figure 82: Delaunay triangulation of land distances with τ = .10 and threshold value method (i). The threshold distance is 17831.40. The map for τ = .25 and threshold value method (i) is identical to this one (with a threshold distance of 14859.50).

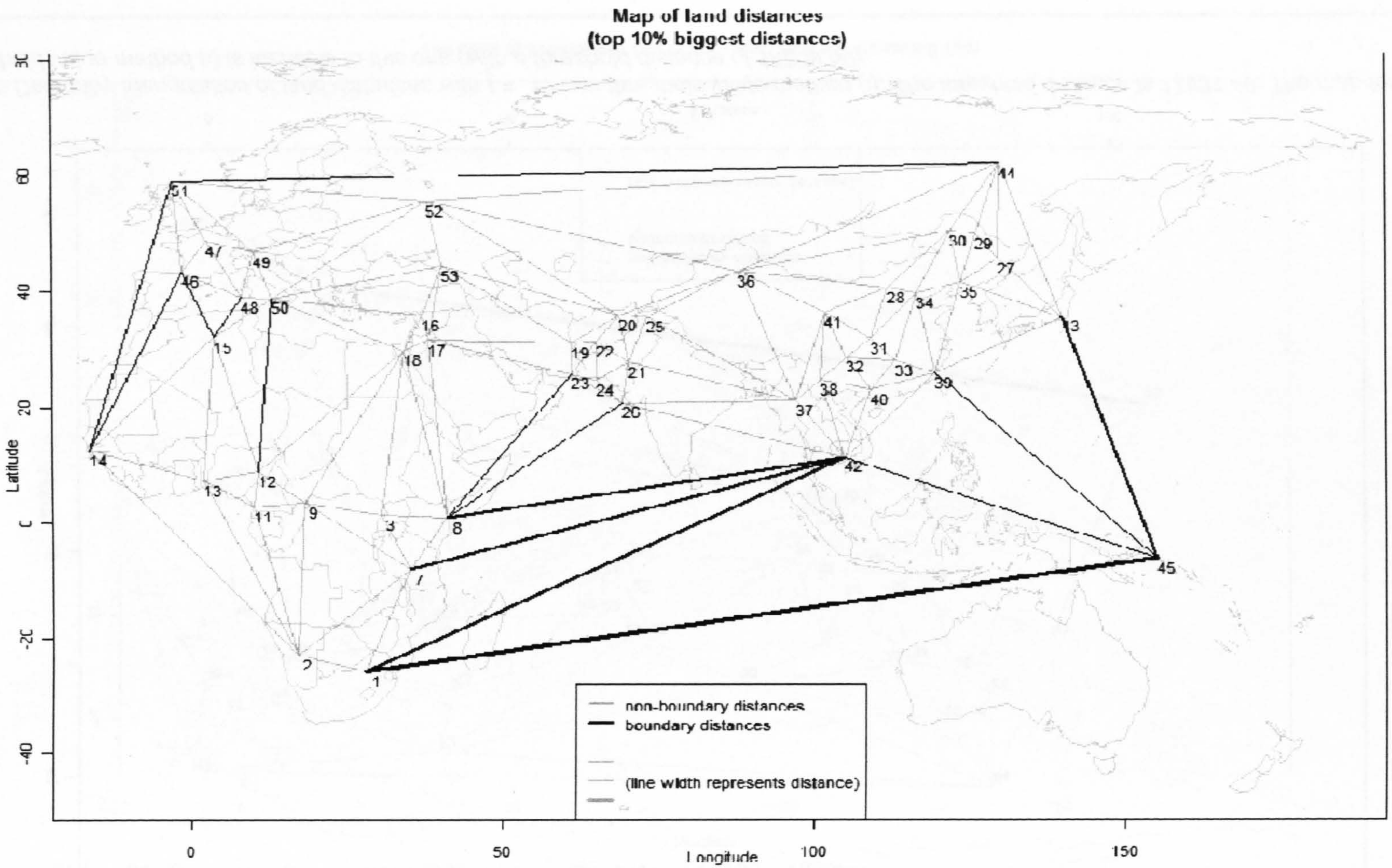

Figure 83: Delaunay triangulation of land distances with τ = .10 and threshold value method (ii). The threshold distance is 5511.11.

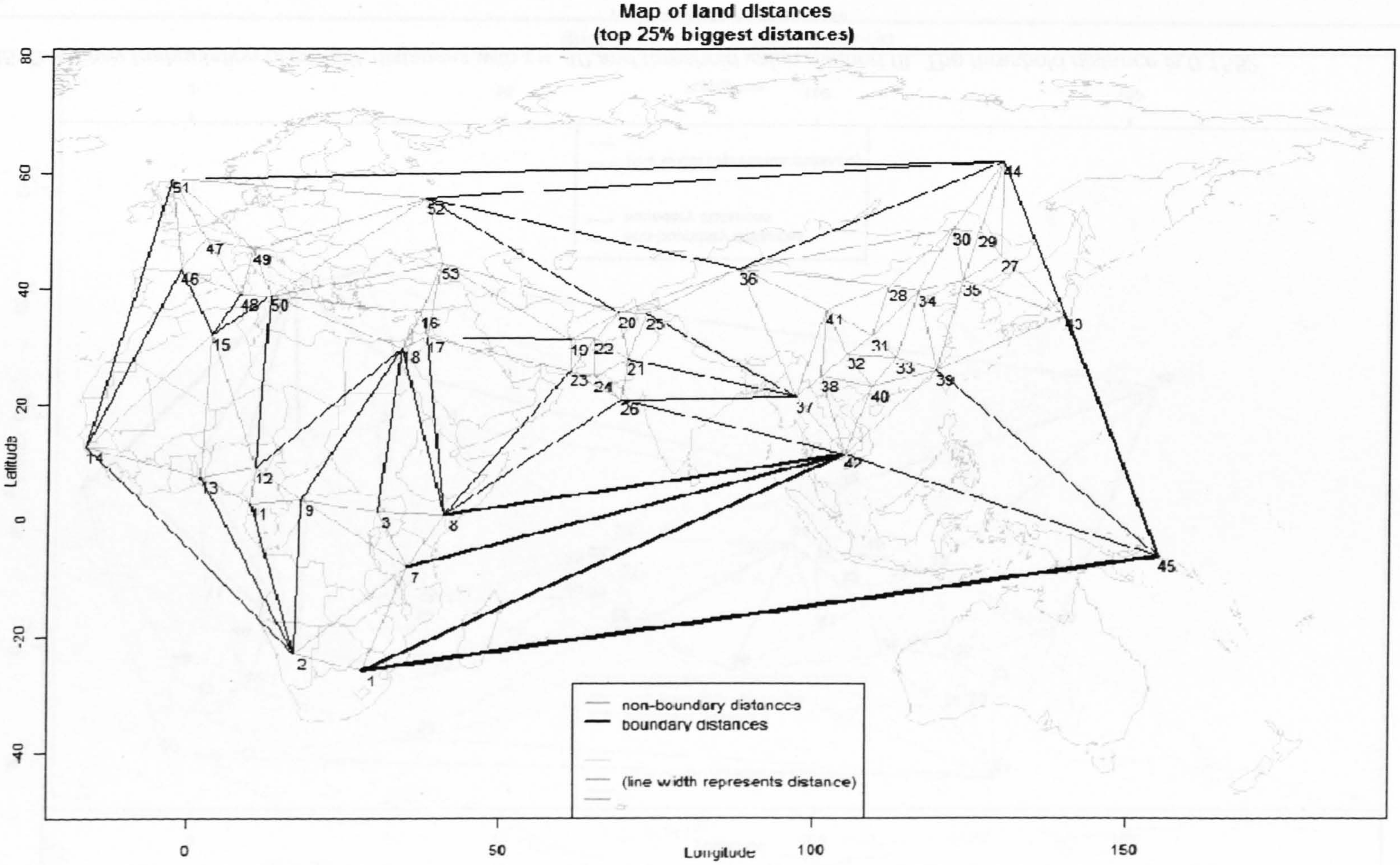

Figure 84: Delaunay triangulation of land distances with τ = .25 and threshold value method (ii). The threshold distance is 2974.62.

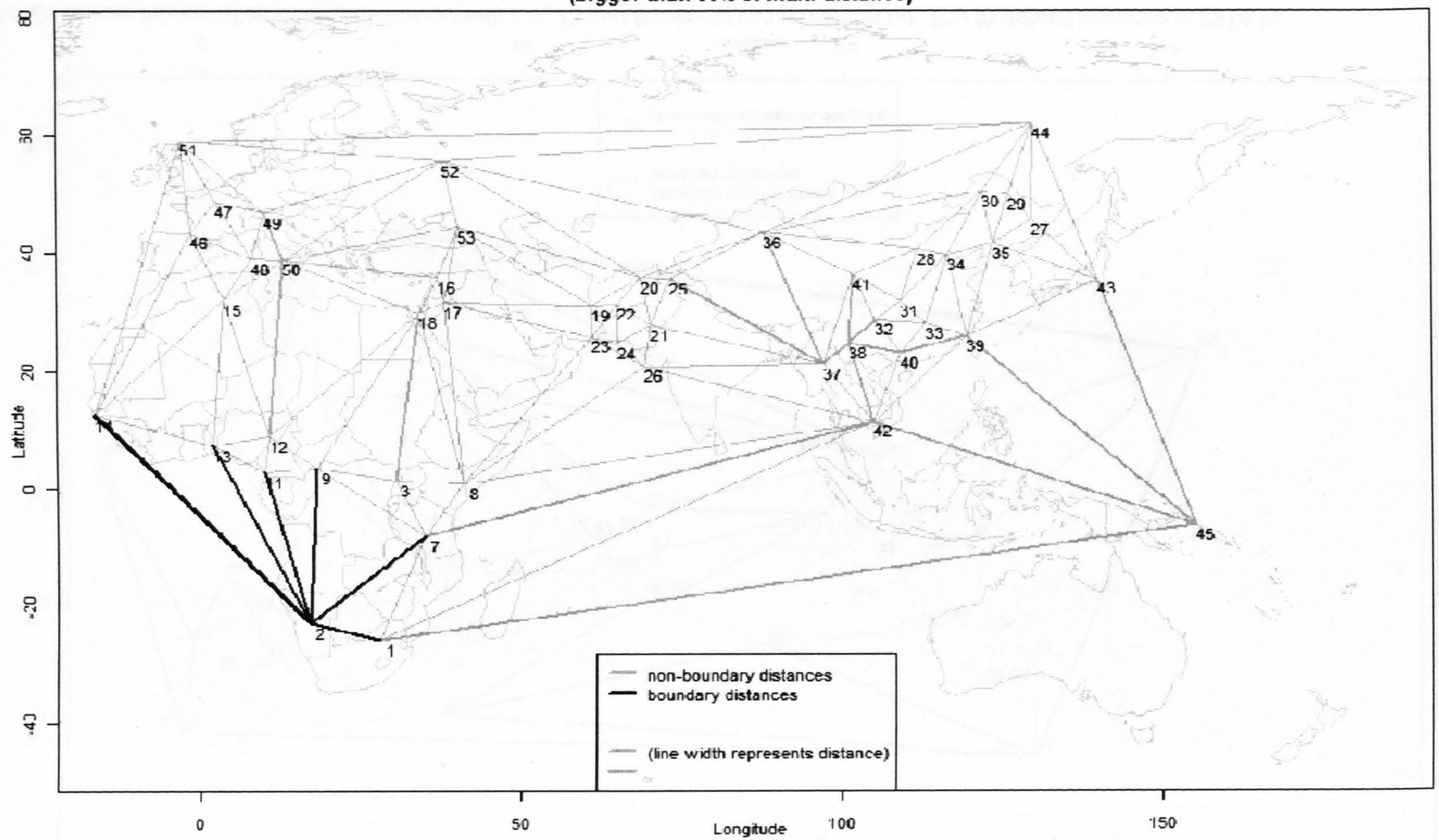

Figure 85: Delaunay triangulation of genetic distances with r = .10 and threshold value method (i). The threshold distance is 0.1552

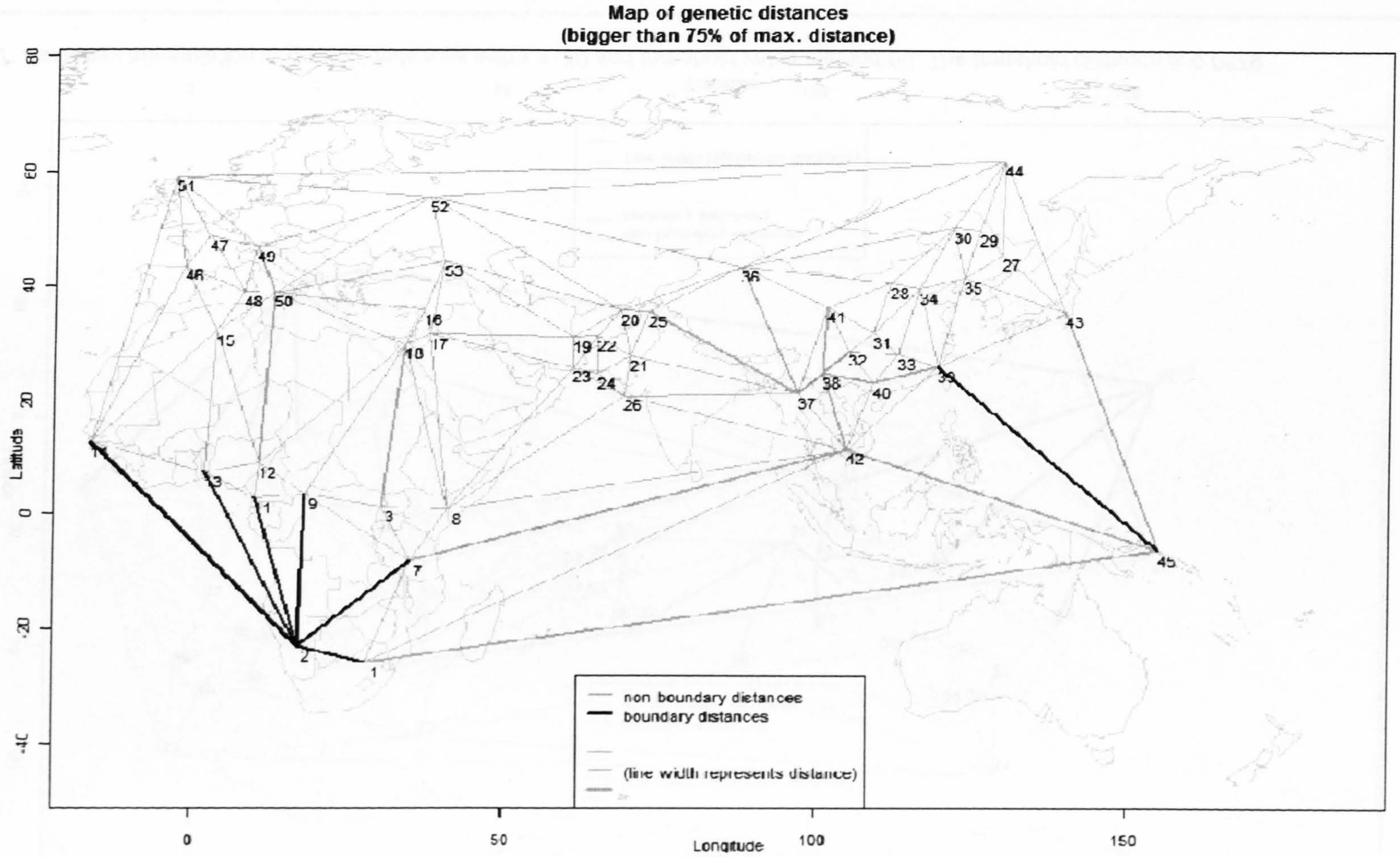

Figure 86: Delaunay triangulation of genetic distances with τ = .25 and threshold value method (i). The threshold distance is 0.0915.

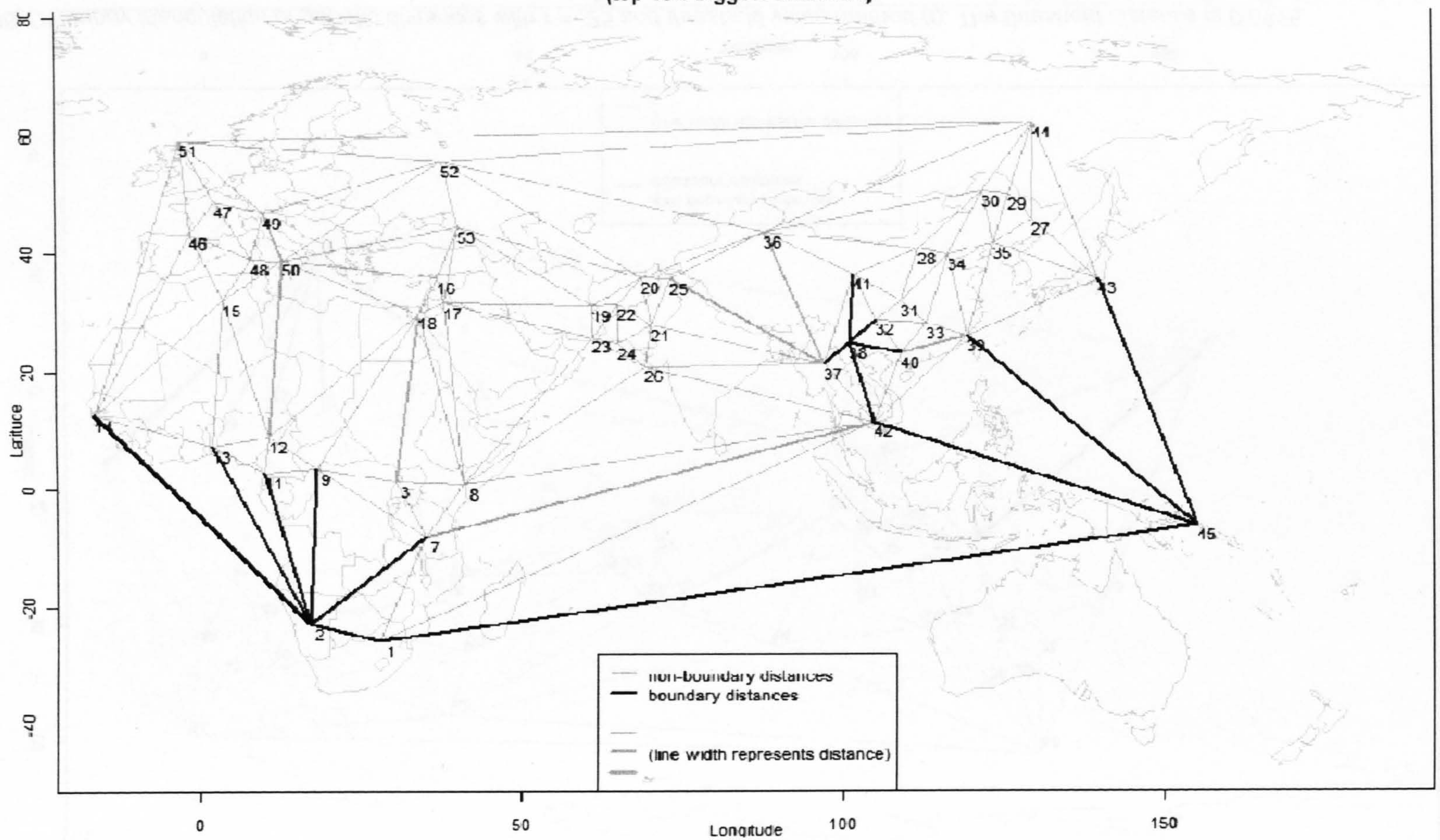

Figure 87: Delaunay triangulation of genetic distances with τ = .10 and threshold value method (ii). The threshold distance is 0.0679.

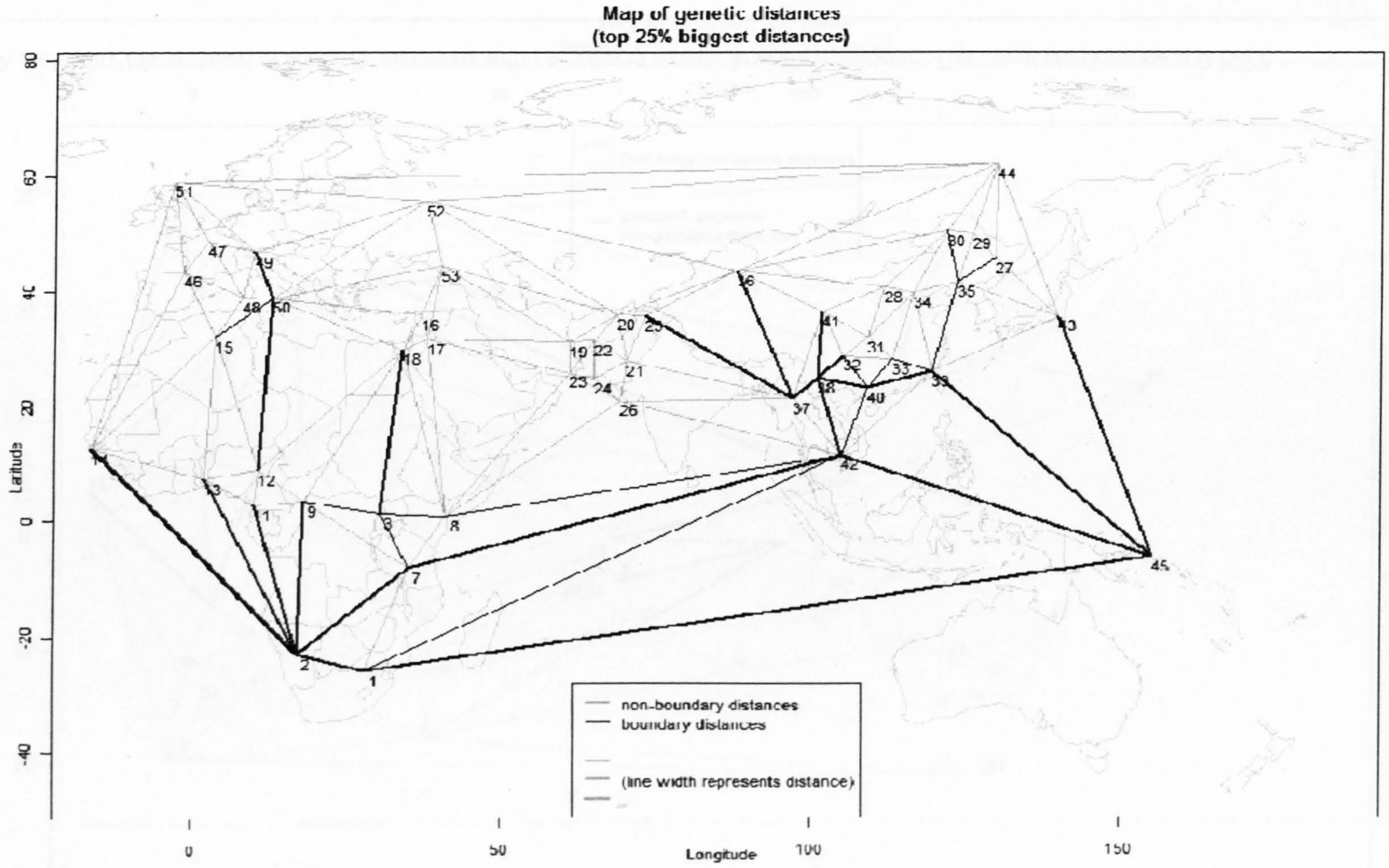

Figure 88: Delaunay triangulation of genetic distances with τ = .25 and threshold value method (ii). The threshold distance is 0.0547.

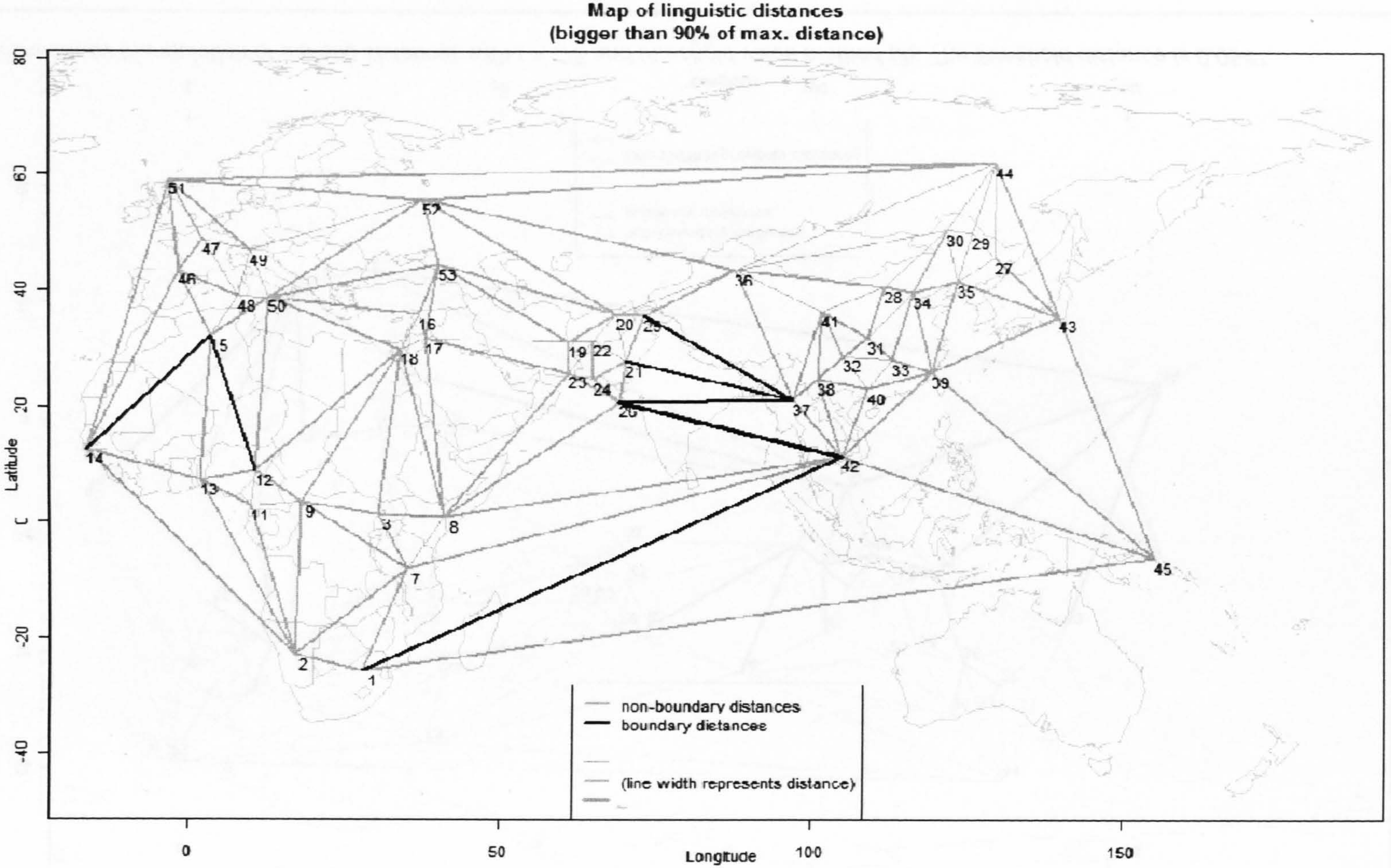

Figure 89: Delaunay triangulation of linguistic distances with τ = .10 and threshold value method (i). The threshold distance is 0.7277.

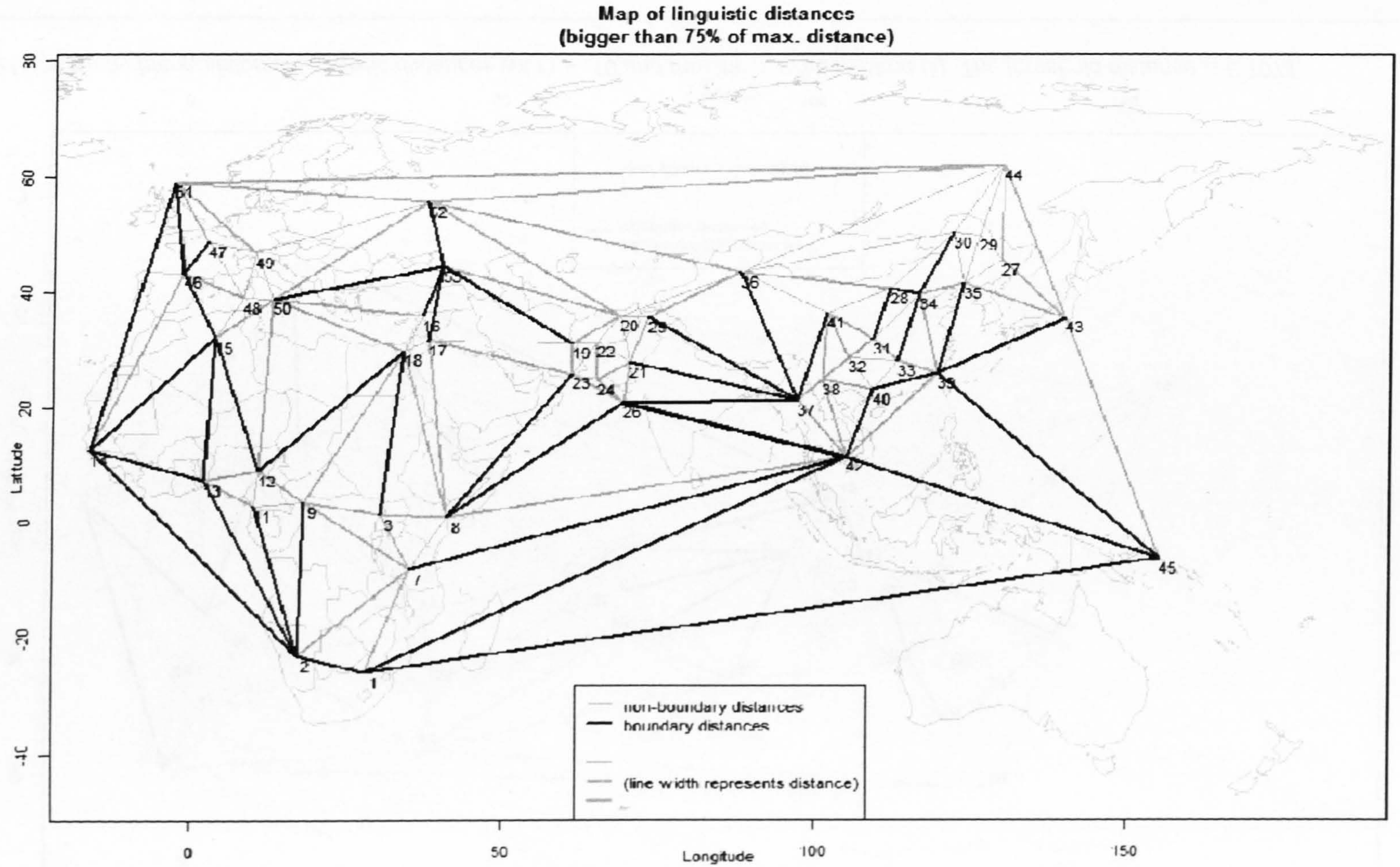

Figure 90: Delaunay triangulation of linguistic distances with τ = .25 and threshold value method (i). The threshold distance is 0.6065.

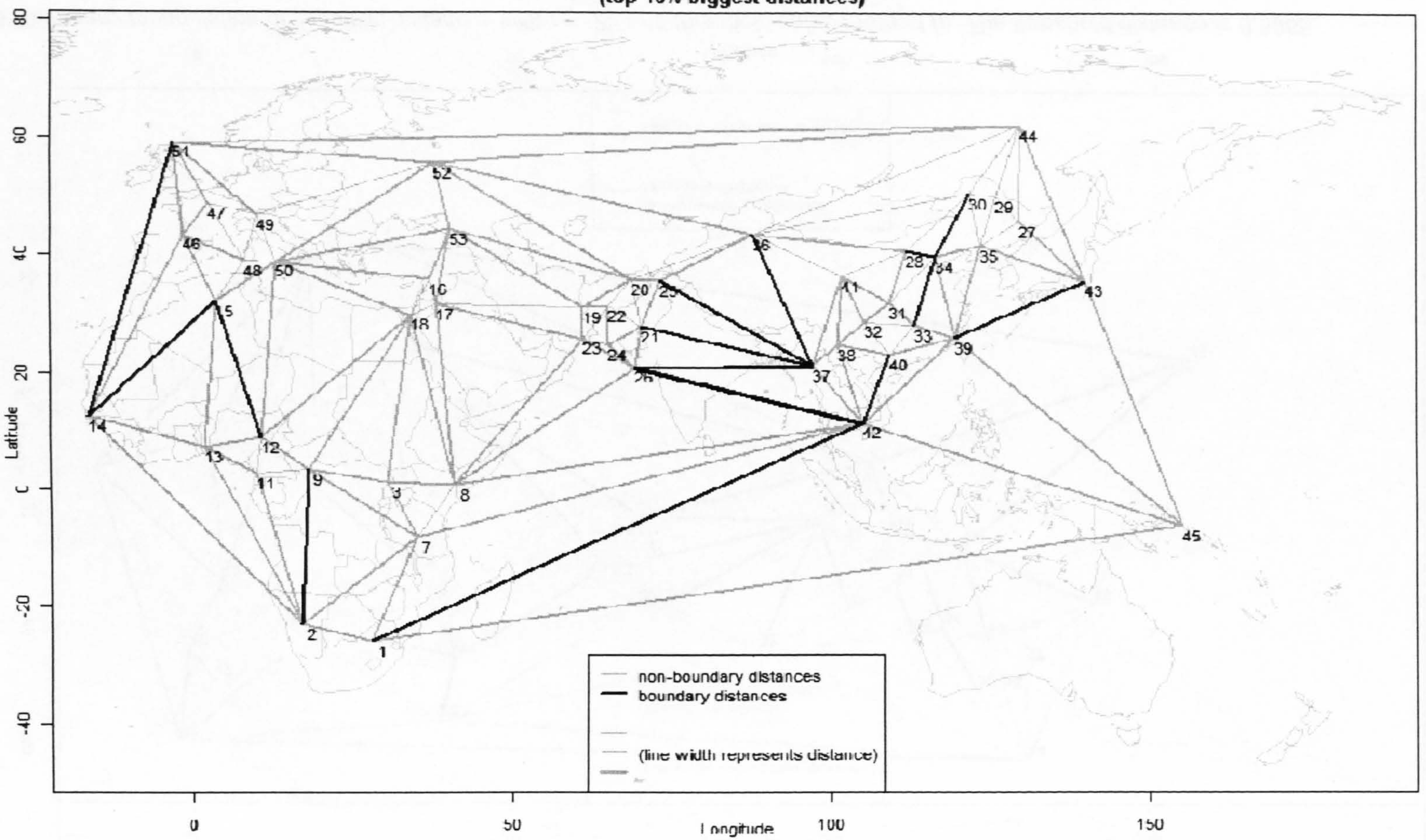

Figure 91: Delaunay triangulation of linguistic distances with τ = .10 and threshold value method (ii). The threshold distance is 0.7071.

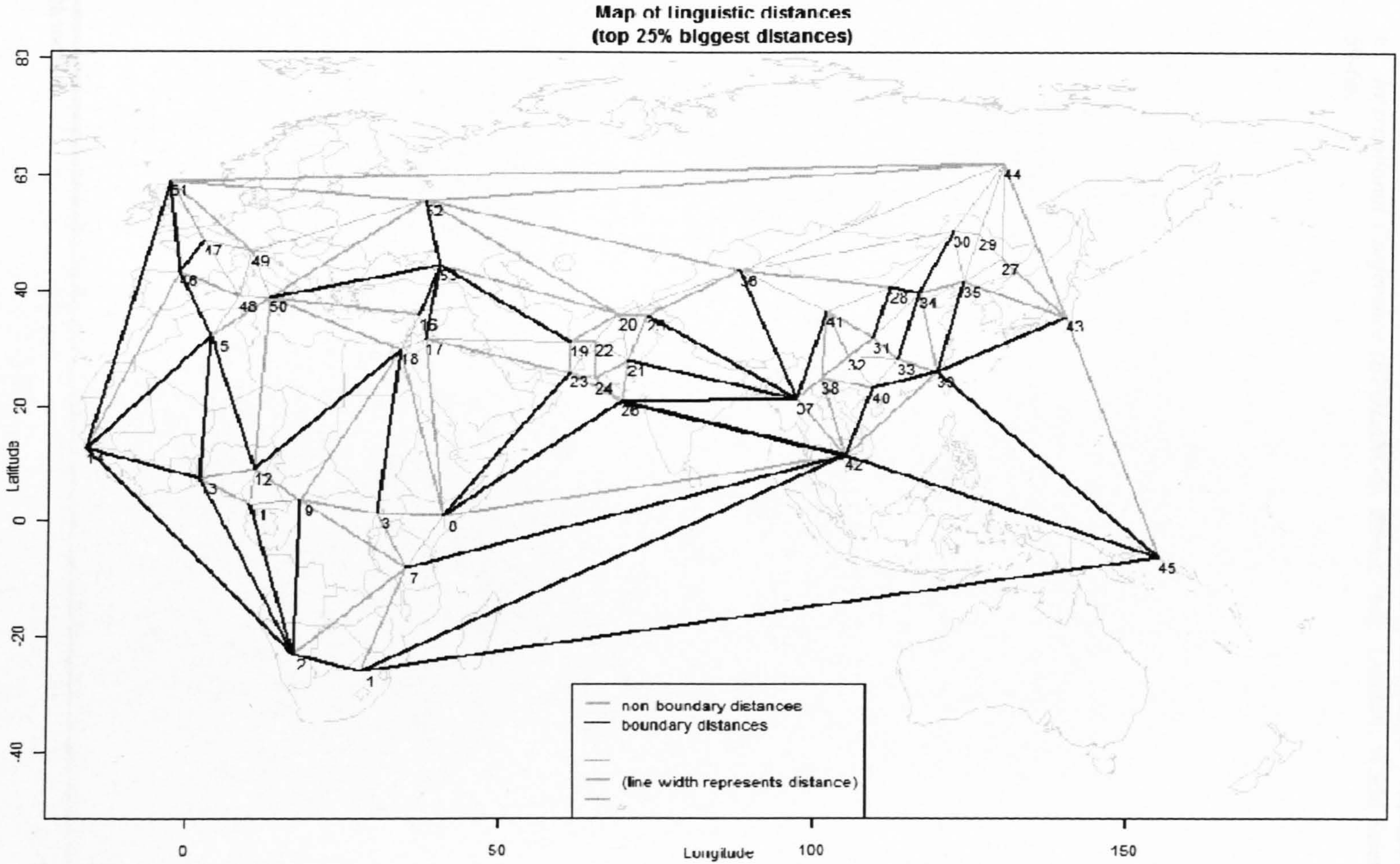

Figure 92: Delaunay triangulation of linguistic distances with τ = .25 and threshold value method (ii). The threshold distance is 0.6202.

Annex 8: Published papers

Annex 8.1: Mostly out of Africa, but what did the others have to say?

DEDIU, D. (2006), *Mostly out of Africa, but what did the others have to say*?, In CANGELOSI, A., SMITH, A. D. M. & SMITH, K. (Eds.), *The Evolution of Language: Proceedings of the 6th International Conference (EVOLANG6), Rome, Italy*. London: World Scientific, pp. 59-66.

Mostly Out Of Africa, But What Did the Others Have To Say?

DAN DEDIU

Language Evolution and Computation Research Unit, University of Edinburgh, 40 George Square, Edinburgh, EH8 9LL, Scotland, UK

The Recent Out-of-Africa human evolutionary model seems to be generally accepted. This impression is very prevalent outside palaeoanthropological circles (including studies of language evolution), but proves to be unwarranted. This paper offers a short review of the main challenges facing ROA and concludes that alternative models based on the concept of metapopulation must be also considered. The implications of such a model for language evolution and diversity are briefly reviewed.

Introduction

As is very well known, the modern human origins debate is now definitely closed and the general consensus is that the Recent Out of Africa model (Stringer & Andrews, 1988) explains perfectly well the genetic, palaeoanthropological and archaeological patterns observed. So, a fairly recent (around 200,000 years ago) and localized (a single population in (East) Africa) origin of modern humans followed by global expansion and replacement explains everything... But, is it really so?

The evidence

The issue of modern human origins is very important, profoundly influencing the range of explanations for the emergence, maintenance and evolution of language and the interactions between population genetic and linguistic structures. The impression outside the palaeoanthropological circles, is that the Recent Out-of-Africa model (henceforth ROA) is true, perception usually reinforced through the popularization press. In fact, there *is* a debate going on and the matters are *very far* from being settled.

I have selected the most recent papers (post 01.2000 but also a few earlier very important ones), dealing with cases where the ROA model does not fit or fits equally well as the alternative models. The search was not exhaustive and the further selection for inclusion in the review was rather strict, but still, the count

is quite large for a "closed" debate. This is the list of the main such points:

The transition to modern *Homo sapiens* was not sudden: the appearance of modern humans is sometimes clad as a heroic myth (McBrearty & Brooks, 2000), as a sudden transition, as a revolution. But there wasn't any such revolution (McBrearty & Brooks, 2000), neither morphologically, nor behaviorally, instead a mosaic of independent transitions to skeletal and behavioral modernity took place in Africa.

The modern humans originated from a structured population: the X chromosome disprove a single panmictic population, favoring models which "incorporate admixture between divergent African branches of the genus *Homo*" (Garrigan *et al.*, 2005a; Harris & Hey, 1999; Harding & McVean, 2004).

Some genes have very deep, non-African branches: the *RRM2P4* pseudogene has a MRC of ~2 MYA in East Asia (Garrigan *et al.*, 2005b), suggesting introgression from archaic local humans. The *dystrophin* gene presents a haplotype predating the ROA expansion and virtually absent from Africa. It might have left Africa earlier and introgressed later (Ziętkiewicz *et al.*, 2003). A noncoding region of the X chromosome (Xq21.1-21.33) shows a variant possibly arisen in Eurasia > 140 KYA (Yu, Fu & Li, 2002). Templeton (2002), applying nested clade analysis, finds a pattern of interbreeding between expanding and local populations.

Regional morphological continuity: one of the oldest claims against ROA-type models (Weidenreich, 1947). Wolpoff *et al.*, (2001) analyzed transitional cranial forms in two peripheral regions (Australia and Czech Republic) and concluded that they have dual ancestry. Wu (2004) concludes evolutionary continuity in China between *sapiens* and *erectus*. Demeter, Manni & Coppens (2003) supports regional continuity in the Far East with a morphometric analysis of 45 fossil crania. The most ancient European modern (Romania) presents a "mosaic of archaic, early modern human and possibly Neandertal morphological features" (Trinkaus *et al.*, 2003). The most well-known such case is the Abrigo do Lagar Velho infantile skeleton (Duarte *et al.*, 1999), showing a mixture of modern and Neanderthal morphological characters (Duarte *et al.*, 1999; Trinkaus & Zilhão, 2003), still accepted despite the critics. Given the burial context, the child was considered as a full community member.

There is also a series of arguments usually considered to support ROA, but which turn out not to be decisive:

Ancient Neanderthal *mtDNA* proves them a different species: the conclusion

from extraction studies (Krings *et al.*, 1997; Lalueza-Fox *et al.*, 2005; Krings *et al.*, 2000; Ovchinnikov *et al.*, 2000) is that Neanderthal *mtDNA* is different from modern, seemingly supporting a replacement model. **But** Gutiérrez, Sánchez & Marín (2002) show ancient *mtDNA* is very sensitive to phylogenetic methods, diagenetic modifications have altered the sequences, and conclude that Neaderthal and modern *mtDNA* may overlap. Nordborg (1998) probabilistically proved that any single locus cannot resolve between replacement and admixture, being necessary to consider many loci in parallel (Wall (2000) suggests 50-100). *mtDNA* was extracted from a fossil modern gracile Australian *Homo sapiens* (Adcock *et al.*, 2001) and proved outside the modern pool. Later, the finds (LM3) were redated to 40±2 KYA (Bowler *et al.*, 2003) and the methodology contested (Cooper *et al.*, 2001), without denying that *mtDNA* lineages can be decoupled from other parts of the genome (Relethford, 2001a).

Based on living primates, the hominid clade was speciose: contested by Hunt (2003), who argues that if appropriate models are considered (the great apes), the hominin lineage may be seen "as a single, phenotypically diverse, reticulately evolving species" (Hunt, 2003).

Neanderthal morphology separates them from moderns: Harvati, Frost & McNulty (2004) used 3D primate craniofacial models and concluded Neanderthals and moderns to be separate species, but Ahern, Hawks & Lee (2005) considered this approach not capable of distinguishing between same or different species. Morphological differences could be due to non-genetic factors (Bogin & Rios, 2003): rapid dramatic morphological changes in modern Mayans accompanies migration to the USA, cautioning against morphological differences in fossil humans as diagnostic for species.

Genetic structure of living populations shows greater diversity in Africa and an African origin of human genes: generally, Africa harbors the greatest genetic diversity of living humans and most gene trees coalesce there (Jobling *et al.*, 2004) but this pattern is not true at least for the X chromosome. The greater genetic diversity of Africa can be explained by a greater long-term population size (Relethford, 2001b), also accommodating the majority coalescence (Takahata, Lee & Satta, 2001).

Modern humans are genetically very uniform: not precluding geographical differentiation (Bamshad *et al.*, 2003) and is usually considered the effect of a major population bottleneck, either a speciation or a migration/founder effect (Jobling *et al.*, 2004) or both. But this can be interpreted as a metapopulation evolutionary history (Relethford, 2001b; Templeton, 2002; Harding & McVean, 2004; Eswaran, 2002), accommodating the small effective population size

(Rousset, 2003) with a large enough adult population. Yu *et al.* (2003) shows the chimpanzees genetic diversity to have been overestimated.

There are some other arguments, like the **relative abundance of hybrids in primates** (Jolly, 2002), suggesting ubiquitous admixture in humans or the **unexpected diversity of our genus**, highlighted by the recent discovery of *Homo floresiensis* (Brown *et al.*, 2004), also pointing to advanced cognitive and technological capacities of *Homo erectus*, allowing him to cross Wallace's line.

The suggested class of alternative models

The data presented above (and more not included) suggests that an alternative class of models should be considered, but choosing it demands awareness to the influence of certain non-scientific factors, like political/moral (Wolpoff & Caspari, 1997), personality clashes/ambitions (Jobling *et al.*, 2004) and favored source (genetic, archaeological, fossil).

Generally, a polarity is described between the ROA model and *multiregionalsim* (Wolpoff & Caspari, 1997; Relethford, 2001b; Lewin, 1998; Jobling *et al.*, 2004), but, (Relethford, 2001b), there are *two* distinct dimensions: the *mode of transition* between archaic and modern humans and the *location and timing of this transition*. Our analysis suggests a recent African origin, a structured ancestral population (metapopulation), a mosaic/accretion of independent traits (morphological and behavioral/cultural) and is disfavoring a speciation event. It suggests a reticulate evolution, where constant gene flow between demes insures local adaptation and continuity while spreading globally the modern genetic-cultural complex. These seem to be satisfied by various models proposed (for example, Relethford, 2001b, Eswaran, 2002 and especially Templeton, 2002), but for our purposes, the following main points are relevant:

- no abrupt speciation event separating moderns from archaics;
- culturally, an accretionary evolution and not a sharp revolution;
- admixture between the migrating waves and locally adapted and differentiated archaics, insuring various degrees of regional continuity;
- metapopulational evolutionary model, whereby demes are constantly created, replaced and extinguished, maintaining genetic and cultural flows, such that there is a global evolutionary accretion of genes and cultural traits without a "core" source population of the full package, Africa being demographically dominant.

Conclusions: implications for language evolution and diversity

Opposed to ROA, such a model can accommodate the language capacity as a mosaic of independent traits evolved in different demes. Language has a more or less specific genetic component, (Stromswold, 2001), confirmed by the *FOXP2* gene (Enard *et al.*, 2002) and seemingly supported by Williams syndrome (Bellugi, Korenberg & Klima, 2001). It is conceivable, for example, that the human-specific *FOXP2* mutations arose in different demes at different times and coalesced with the qualitatively different languages they allowed. The discovery (Mekel-Bobrov *et al.*, 2005; Evans *et al.*, 2005) of recent variants of two genes related to brain growth and development, with signatures of strong positive natural selection, not yet fixated and with marked population structures supports this mosaic evolutionary process.

There could exist minor inter-populational genetic differences in linguistic capacity (because of regional continuity, founder effect or not yet fixated advantageous alleles), offering new perspectives on language evolution, given that the basic requirement is heritable variation. Such a model highlights the early evolution of the language capacity and languages as two inter-related phenomena in metapopulations, leading to the modern linguistic capacity, able to support an immense linguistic (almost neutral) variation.

Another possibility is that besides the accidental correlations between genes and languages (Cavalli-Sforza *et al.*, 1994), there might also exist a slight non-accidental correlation, whereby specific genetic configurations favor/are favored by specific linguistic features. A fictional example could be a population with a high incidence of articulatory incapacity to produce a trilled /r/, which in turn will select for languages realizing the phoneme /r/ as an approximant. Conversely, speakers with such a deficiency will not incur any fitness penalty when immersed into a community speaking the /r/-approximant language. This hypothetical example can be extended to more plausible cases, like the better control of rapid orofacial movements (supposedly) brought by the human-specific mutation(s) in *FOXP2*.

Acknowledgements

My work has been supported by an ORS Awards Grant 2003014001 and a CHSS Studentship, The University of Edinburgh.

References

Adcock, G. J., Dennis, E. S., Easteal, S., Huttley, G. A., Jermiin, L. S., Peacock, W. J. & Thorne, A. (2001). Mitochondrial DNA sequences in ancient Australians. *PNAS*, 98, 537-542.

Ahern, J. C. M., Hawks, J. D. & Lee, S.-H. (2005). Neandertal taxonomy reconsidered... again. *Journal of Human Evolution*, 48, 647-652.

Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A.& Jorde, L. B. (2003). Human population genetic structure and inference of group membership. *American Journal of Human Genetics*, 72, 578-589.

Bellugi, U., Korenberg, J. R. & Klima, E. S. (2001). Williams syndrome. *Clinical Neuroscience Research*, 1, 217-229.

Bogin, B. & Rios, L. (2003). Rapid morphological change in living humans. *Comparative Biochemistry and Physiology Part A*, 136, 71-84.

Bowler, J. M., Johnston, H., Olley, J. M., Prescott, J. R., Roberts, R. G., Shawcross, W. & Spooner, N. A. (2003). New ages for human ocuppation and climatic change at Lake Mungo, Australia. *Nature*, 421, 837-840.

Brown, P., Sutikna, T., Morwood, M. J., Soejono, Jatmiko, Saptomo, E. W. & Due, R. A. (2004). A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature*, 431, 1055-1061

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Cooper, A., Rambaut, A., Macaulay, V., Willerslev, E., Hansen, A. J. & Stringer, C. (2001). Human origins and ancient human DNA. *Science*, 292, 1655-1656.

Demeter, F., Manni, F. & Coppens, Y. (2003). Late Upper Pleistocene human peopling of the Far East. *Comptes Rendus Palevol*, 2, 625-638.

Duarte, C., Maurício, J., Pettitt, P. B., Souto, P., Trinkaus, E., van der Plicht, H. & Zilhão, J. (1999). The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *PNAS*, 96, 7604-7609.

Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S. L., Wiebe, V., Kitano, *et al.* (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, 418, 869-872.

Eswaran, V. (2002). A diffusion wave out of Africa. *Current Anthropology*, 43, 749-774.

Evans, P. D., Gilbert, S. L., Mekel-Bobrov, N., Vallender, E. J., Anderson, Vaez-Azizi, L. M., *et al.* (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science*, 309, 1717-1720.

Garrigan, D., Mobasher, Z., Kingan, S. B., Wilder, J. A. & Hammer, M. F. (2005a). Deep haplotype devrgence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics*, 170, 1849-1856.

Garrigan, D., Mobasher, Z., Severson, T., Wilder, J. A. & Hammer, M. F. (2005b). Evidence for archaic Asian ancestry on the human X chromosome. *Molecular Biology and Evolution*, 22, 189-192.

Gutiérrez, G., Sánchez, D. & Marín, A. (2002). A reanalysis of the ancient mitochondrial DNA sequences recovered from Neanderthal bones. *Molecular Biology and Evolution*, 19, 1359-1366.

Harding, R. M. & McVean, G. (2004). A structured ancestral population for the evolution of modern humans. *Current Opinion in Genetics & Development*, 14, 667-674.

Harris, E. E. & Hey, J. (1999). X chromosome evidence for ancient human histories. *PNAS*, 96, 3320-3324.

Harvati, K., Frost, S. R. & McNulty, K. P. (2004). Neanderthal taxonomy reconsidered: implications of 3D primate models of intra- and interspecific differences. *PNAS*, 101, 1147-1152.

Hunt, K. D. (2003). The Single Species Hypothesis. *Human Biology*, 75, 485-502.

Jobling, M. A., Hurles, M. E. & Tyler-Smith, C. (2004). *Human evolutionary genetics: origins, peoples and disease*. Garland Science.

Jolly, C. J. (2002). A proper study for mankind: analogies from the papionin monkeys and their implications for human evolution. *American Journal of Physical Anthropology*, 116, 177-204.

Krings, M., Capelli, C., Tschentcher, F., Geisert, H., Meyer, S., von Haeseler, *et al.* (2000). A view of Neandertal genetic diversity. *Nature Genetics*, 26, 144-146.

Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking. M. & Pääbo, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell*, 90, 19-30.

Lalueza-Fox, C., Lourdes Sampietro, M., Caramelli, D., Puder, Y., Lari. M., Calafell, *et al.* (2005). Neandertal evolutionary genetics. *Molecular Biology and Evolution*, 22, 1077-1081.

Lewin, R. (1998). *Principles of Human Evolution.* Blackwell Science.

Long, J. C. & Kittles, R. A. (2003). Human genetic diversity and the nonexistence of biological rances. *Human Biology*, 75, 449-471.

McBrearty, S., & Brooks, A. S. (2000). The revolution that wasn't. *Journal of Human Evolution*, 39, 453-563.

Mekel-Bobrov, N., Gilbert, S. L., Evans, P. D., Vallender, E. J., Anderson, J., R., Hudson, *et al.* (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. *Science*, 309, 1720-1722.

Nordborg, M.. (1998). On the probability of Neanderthal ancestry. *American*

Journal of Human Genetics, 63, 1237-1240.
Ovchinnikov, I. V., Götherström, A., Romanova, G. P., Kharitonov, V. M., Lidén. K. & Goodwin, W. (2000). Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, 404, 490-493.
Relethford, J. H. (2001a). Ancient DNA and the origins of modern humans. *PNAS*, 98, 390-391.
Relethford, J. H. (2001b). *Genetics and the search for modern human origins*. Wiley-Liss.
Rousset, F. (2003). Effective size in simple metapopulation models. *Heredity*, 91, 107-111.
Stringer, C. & Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science*, 239, 1263-1268.
Stromswold, K. (2001). The heritability of language. *Language*, 77, 647-723.
Takahata, N., Lee, S.-H. & Satta, Y. (2001). Testing multiregionality of modern human origins. *Molecular Biology and Evolution*, 18, 172-183.
Templeton, A. R. (2002). Out of Africa again and again. *Nature*, 416, 45-51.
Trinkaus, E. & Zilhão, J. (2003). Phylogenetic implications. In J. Zilhão and E. Trinkaus (Eds.), *Portrait of the Artist As a Child.* (pp. 497-518). Oxbow Books Ltd.
Trinkaus, E., Moldovan, O., Milota, Ş., Bîlgăr, A., Sarcina, L., Athreya, S., *et al.* (2003). An early modern human from the Peştera cu Oase, Romania. *PNAS*, 100, 11231-11236.
Wall, J. D. (2000). Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, 154, 1271-1279.
Weidenreich, F. (1947). Facts and speculations concerning the origin of Homo sapiens. *American Anthropologist*, 49, 187-203.
Wolpoff, M. H. & Caspari, R. (1997). *Race and human evolution*. Westview Press.
Wolpoff, M. H., Hawks, J., Frayer, D. W. & Hunley, K. (2001). Modern human ancestry at the peripheries. *Science*, 291, 293-297.
Wu, X. (2004). On the origin of modern humans in China. *Quaternary International*, 117, 131-140.
Yu, N., Fu, Y.-X.& Li, W.-H. (2002). DNA polymorphism in a worldwide sample of human X chromosomes. *Molecular Biology and Evolution*, 19, 2131-2141.
Yu, N., Jenesen-Seaman, M. I., Chemnick, L., Kidd, J. R., Deinard, A. S., Ryder, O., Kidd, *et al.* (2003). Low nucleotide diversity in chimpanzees and bonobos. *Genetics*, 164, 1511-1518.
Ziętkiewicz, E., Yotova, V., Gehl, D., Wambach, T., Arrieta, I., Batzer, M., *et al.* (2003). Haplotypes in the Dystrophin DNA segment point to a mosaic origin of modern human diversity. *American Journal of Human Genetics*, 73, 994-1015.

References

ADCOCK, G. J., DENNIS, E. S., EASTEAL, S., HUTTLEY, G. A., JERMIIN, L. S, PEACOCK, W. J. & THORNE, A. (2001a). Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *PNAS*, 98, 537-542.

ADCOCK, G. J., DENNIS, E. S., EASTEAL, S., HUTTLEY, G. A., JERMIIN, L. S, PEACOCK, W. J. & THORNE, A. (2001b). Response to "Human origins and ancient human DNA". *Science*, 292, 1656-1656.

AGUIRRE, E. & CARBONELL, E. (2001), Early human expansions into Eurasia: The Atapuerca evidence, *Quaternary International* 75:11-18

AIKHENVALD, A. Y. & DIXON, R. M. W. (Eds.) (2001), *Areal diffusion and genetic inheritance*, Oxford University Press: NY

ALFRED Database: The Allele Frequency Database, http://alfred.med.yale.edu/alfred/index.asp

ALLMAN, E. S. & RHODES, J. A. (2004), *Mathematical Models in Biology: An Introduction*, Cambridge University Press: Cambridge, UK

AMMERMAN, A. J. & CAVALLI-SFORZA, L. L. (1984), *The Neolithic transition and the genetics of populations in Europe*, Princeton University Press: NJ

ANTÓN, S. C. (2003), Natural History of Homo erectus, *Yearbook of Physical Anthropology* 46:126–170

APPLEYARD, D. (1999), *Afroasiatic and the Nostratic hypothesis*, In RENFREW, C. & NETTLE, D. (Eds.) (1999), pp.: 289-325

ARENSBURG, B. & TILLIER, A.M. (1991), Speech and the Neanderthals, *Endeavour* 15:26-8

ARENSBURG, B., TILLIER, A. M., VANDERMEERSCH, B., DUDAY, H., SCHEPARTZ, L.A. &RAK, Y. (1989) A Middle Palaeolithic human hyoid bone, *Nature* 338:758–760

ARGUE, D., DONLON, D., GROVES, C. & WRIGHT, R. (in press), *Homo floresiensis*: microcephalic, pygmoid, *Australopithecus*, or *Homo*?, *Journal of Human Evolution*.

ARNAIZ-VILLENA, A., MARTIŃEZ-LASO, J. & ALONSO-GARCIÁ, J. (2001), The correlation between languages and genes: The Usko-Mediterranean peoples, *Human Immunology* 62:1051-1061

ARNOLD, M. L., KENTNER, E. K., JOHNSTON, J. A., CORNMAN, S. & BOUCK, A. C. (2001), Natural hybridization and fitness, *Taxon* 50:93-104.

ASFAW, B., GILBERT, W. H., BEYENE, Y., HART, W. K., RENNE, P. R., GABRIEL, G. W., VRBA, E. S. & WHITE, T. D. (2002) Remains of Homo erectus from Bouri, Middle Awash, Ethiopia, *Nature* 416:317-320

ASIMOV, I. & SILVERBERG, R. (1993), *The Ugly Little Boy*, Bantam Books

ATRAN, S. (1998), Folk biology and the anthropology of science: cognitive universals and cultural particulars, *Behavioural and Brain Sciences* 21:547-569

AWADALLA, P. (2004), *Opinion: Does mtDNA recombine*. In Jobling, Hurles & Tyler-Smith (2004):42-43 (Box 2.9)

BAMSHAD, M. J., WOODING, S., WATKINS, W. S., OSTLER, C. T., BATZER, M. A. & JORDE, L. B.

(2003), Human Population Genetic Structure and Inference of Group Membership, *American Journal of Human Genetics* 72:578–589

BANDELT, H.-J., MACAULAY, V. & RICHARDS, M. (2002), *What Molecules Can't Tell Us about the Spread of Languages and the Neolithic*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:99-107

BANTON, M. (1998), *Racial Theories* (2nd Edition), Cambridge: Cambridge University Press

BAR-YOSEF, O. (2002), *The Natufian Culture and the Early Neolithic: Social and Economic Trends in Southwestern Asia*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:113-126

BARBUJANI, G. & DUPANLOUP, I. (2002), *DNA Variation in Europe: Estimating the Demographic Impact of Neolithic Dispersals*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:421-433

BARBUJANI, G. & PILASTRO, A. (1993), Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily, *Proceedings and the National Academy of Sciences of the USA* 90:4670-4673

BARBUJANI, G., BERTORELLE, G. & CHIKHI, L. (1998), Evidence for Palaeolithic and Neolithic gene flow in Europe, *American Journal of Human Genetics* 62:488-491

BARKER, G. (2002), *Transitions to farming and pastoralism in North Africa*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:151-162

BASHIR, E. L. (1991), *A Contrastive Analysis of Brahui and Urdu.* Washington DC: Academy for Educational Development.

BATEMAN, R., GODDARD, I., O'GRADY, R., FUNK, V. A., MOOI, R., KRESS, W. J. & CANNELL, P. (1990), Speaking of Forked Tongues: The feasibility of reconciling human phylogeny and the history of language, *Current Anthropology* 31:1-13

BAXTER, S. (2003), *Evolution*, London: Gollancz

BEAUMONT, P., VILLIERS, H. DE; VOGEL, J. C. (1978), Modern man in Sub-Saharan Africa prior to 49 000 years B.P.: A review and evaluation with particular reference to Border Cave, *South African Journal of Science* (Pretoria) 74:409-419.

BEGUN, D. R. (2004), The Earliest Hominins—Is Less More?, *Science* 303:1478-1480

BELLWOOD, P. (2002), *Farmers, Foragers, Languages, Genes: the Genesis of Agricultural Societies*, In BELLWOOD & RENFREW (Eds.) (2002), pp:17-28

BELLWOOD, P. & RENFREW, C. (Eds.) (2002), *Examining the farming/language dispersal hypothesis*, McDonald Institute for Archaeological Research: Cambridge, UK

BENEDETTO, D., CAGLIOTI, E. & LORETO, V. (2002), Language trees and zipping, *Physical Review Letters* 88, 048702

BENGTSON, J. D. (1998), Editorial: What is Nostratic?, *Mother Tongue* 31:33-38

BENGTSON, J.D. & RUHLEN, M. (1994) *Global etymologies.* In *On the Origin of Languages: Studies in Linguistic Taxonomy* (RUHLEN, M., Ed.), pp. 277–336, Stanford University Press

BERLOCHER, S. (1998), *A brief history of research on speciation*, In HOWARD & BERLOCHER (Eds), 1998:3-15

BICKERTON, D. (2000), *How protolanguage became language*, In KNIGHT, STUDDERT-KENNEDY & HURFORD (Eds.)

BICKERTON, D. (2002), *From protolanguage to language*, In CROW (2002):103-120

BICKERTON, D. (*in press*), Language evolution: a brief guide for linguists, *Lingua* (2005)

BISHOP, D. V. M. (2003). Genetic and environmental risks for specific language impairment in children, *International Journal of Pediatric Otorhinolaryngology* 6751:S143-S157

BLUST, R. (2000), *Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages*, In RENFREW, MCMAHON & TRASK (2000), pp 311-331

BOBE, R., BEHRENSMEYER, A. K. (2004) The expansion of grassland ecosystems in Africa in relation to mammalian evolution and the origin of the genus *Homo*, *Palaeogeography, Palaeoclimatology, Palaeoecology* 207:399-420

BOCHERENS, H., DRUCKER, D. G., BILLIOU, D., PATOU-MATHIS, M. & VANDERMEERSCH, B. (2005), Isotopic evidence for diet and subsistence pattern of the Saint-Césaire I Neanderthal: review and use of a multi-source mixing model, *Journal of Human Evolution* 49:71-87

BOLNICK, D. A., SHOOK, B. A., CAMPBELL, L. & GODDARD, I. (2004), Problematic use of Greenberg's linguistic classification of the Americas in studies of Native American genetic variation, *American Journal of Human Genetics* 75:519-523

BOMHARD, A. R. (1998), *Nostratic, Eurasiatic and Indo-European*, In SALMONS & JOSEPH (Eds.), pp.:17-39

BOMHARD, A. R. (1999), *Review of Dolgopolsky's The Nostratic Macrofamily and Linguistic Palaeontology*, In RENFREW, C. & NETTLE, D. (Eds.) (1999), pp.: 47-74

BONNEAU, D., VERNY, C. & UZÉ, J. (2004), Les facteurs génétiques dans les troubles spécifiques du langage oral, *Archives de pédiatrie* 10:1213-1216.

BONNET, E. & VAN DE PEER, Y. (2002), zt: a software tool for simple and partial Mantel tests, *Journal of Statistical software* 7:1-12.

BOWLER, J. M., JOHNSTON, H., OLLEY, J. M., PRESCOTT, J. R., ROBERTS, R. G., SHAWCROSS, W. & SPOONER, N. A. (2003). New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature*, 421, 837-840.

BRADMAN, N., THOMAS, M. G., WEALE, M. E. & GOLDSTEIN, D. B. (2004), *The Lemba: the 'Black Jews' of Southern Africa*, In JOBLING, HURLES & TYLER-SMITH (2004), p.391, Box 12.2

BRÄUER, G. (1984), A craniological approach to the origin of anatomically modern *Homo sapiens* in Africa and implications for the appearance of modern Europeans. In SMITH, F. H.; SPENCER. F., (eds.), *The Origins of Modern Humans*. New York: Alan R. Liss, p. 327-410.

BRAUNER, S. (1974). *Lehrbuch des Bambara*. Leipzig: VEB Verlag Enzyklopadie.

BRIGHTON, H. (2003), *Simplicity as a driving force in linguistic evolution*, PhD. Thesis, The University of Edinburgh (http://141.14.165.6/users/brighton/brighton_brighton_phd_thesis.ps, August 2006)

BROWN, P., SUTIKNA, T., MOROWOOD, M. J., SOEJONO, R. P., JATMIKO, SAPTOMO, E. W. & ROKUS AWE DUE (2004), A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia, *Nature* 431:1055-1061

BRUCE, G. (2004), *An intonational typology of Swdish*, In BEL, B. & MARLIEN, I. (Eds), *Speech Prosody 2004, International Conference*; Nara, Japan, March 23-26, 2004, ISCA Archive, http://www.isca-speech.org/archive/sp2004, pp. 175-178. Online

http://www.isca-speech.org/archive/sp2004/sp04_175.pdf, August 2006.

Bruce, H. A. & Margolis, R. L. (2002), *FOXP2*: novel exons, splice variants, and CAG repeat length stability. *Human Genetics.* 111:136-144.

Brumm, A., Aziz, F., van den Bergh, G. D., Morwood, M. J., Moore, M. W., Kurniawan, I., Hobbs, D. R. & Fullagar, R. (2006), Early stone technology on Flores and its implications for *Homo floresiensis*, *Nature* 441:624-628.

Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J.-R., De Bonis, L., Coppens, Y., Dejax, J., Denys, C., Duringer, P., Eisenmann, V., Fanone, G., Fronty, P., Geraads, D., Lehmann, T., Lihoreau, F., Louchart, A., Mahamat, A., Merceron, G., Mouchelin, G., Otero, O., Campomanes, P. P., De Leon, M. P., Rage, J.-C., Sapanet, M., Schuster, M., Sudre, J., Tassy, P., Valentin, X., Vignaud, P., Viriot, L., Zazzo, A. & Zollikofer, C. (2002), A new hominid from the Upper Miocene of Chad, Central Africa, *Nature* 418:145-151

Bull, V. J. (2003), *Genealogy and Speciation in Heliconius Butterflies*, PhD Thesis, Dept. of Biology, University College London, June 2003.

Bye, P. (2004), *Evolutionary typology and Scandinavian pitch accent*, Ms. University of Tromsø. Online http://www.hum.uit.no/a/bye/Papers/pitch-accent-kluw.pdf (Ausgust 2006).

Cameron, D.W. (2003), Early hominin speciation at the Plio/Pleistocene transition, *HOMO - Journal of Comparative Human Biology* 54:1-28

Campbell, G. L. (2000) *Compendium of the World's Languages* (Vols. 1 & 2), Second Edition, London: Routledge

Campbell, L. (1999), *Nostratic and linguistic palaeontology in methodological perspective*, In Renfrew, C. & Nettle, D. (Eds.) (1999), pp.: 179-230

Campbell, L. (2002), *Why and How do languages Diversify and Spread*? In Bellwood & Renfrew (Eds.) (2002), pp:49-63

Campbell, L. (2004), *Historical linguistics : an introduction*, Edinburgh: Edinburgh University Press

Cangelosi, A., Smith, A.D.M. & Smith, K. (Eds.) (2006), *The Evolution of Language – Proceedings of the 6th International Conference (EVOLANG6)*, World Scientific, Singapore

Cann, R., Stoneking, M. and Wilson, A. (1987), Mitochondrial DNA and human evolution, *Nature* 325:31-36

car R package: John Fox (2006). *car: Companion to Applied Regression*. R package version 1.1-0. http://www.r-project.org, http://socserv.socsci.mcmaster.ca/jfox/

Caramelli, D., Lalueza-Fox, C., Vernesi, M., Casoli, A., Mallegni, F., Chiarelli, B., Dupanloup, I., Bertanpetit, J., Barbujani, G. & Bertorelle, G. (2003), Evidence for a genetic discontinuity between Neadertals and 24,000-year-old anatomically modern Europeans, *PNAS* 100:6593-6597

Cavalli-Sforza, L. L. (2000), *Genes, Peoples, and Languages*, Allen Lane - The Penguin Press, London

Cavalli-Sforza, L. L. (2002), Demic diffusion ans the basic process of human expansions, In Bellwood & Renfrew (Eds.) (2002), pp.:79-88

CAVALLI-SFORZA, L. L., MENOZZI, P. & PIAZZA, A. (1994), *The history and geography of human genes* (Abridged paperback edition), Princeton: Princeton University Press

CAVALLI-SFORZA, L. L., PIAZZA, A., MENOZZI, P. & MOUNTAIN, J. (1988), Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data, *Proceedings of the National Academy of Sciences of the USA* 85:6002-6006

CAVALLI-SFORZA, L. L., PIAZZA, A., MENOZZI, P. & MOUNTAIN, J. (1989), Genetic and linguistic evolution, *Science* 244:1128-1129

CAVALLI-SFORZA, L.L., MENOZZI, P. & PIAZZA, A. (1993), Demic expansions and human evolution, *Science* 259:639-646.

CHIKHI, L. (2002), *Admixture and the Demic Diffusion Model in Europe*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:435-447

CHOMSKY, N. (1965), *Aspects of a theory of syntax*, MIT Press

CHRISTIANSEN, M.H. & KIRBY, S. (Eds.) (2003), *Language Evolution*, Oxford University Press

CIORĂNESCU, A. (2002), *Dicţionarul Etimologic Al Limbii Române*, Editura Saeculum I.O.: Bucureşti

COHEN, M. N. (2002), *The Economies of Late Pre-farming and Farming Communities and their Relation to the Problem of Dispersals*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:41-47

COMRIE, B. (1981), *Language Universals and Linguistic Typology*, Basil Blackwell Publisher Limited: Oxford, UK

COMRIE, B. (2002), *Farming dispersal in Europe and the spread of the Indo-European language family*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:409-419

COON, C.S. (1962), *The Origin of Races*. New York: Knopf.

COOPER, A., RAMBAUT, A., MACAULAY, V., WILLERSLEV, E., HANSEN, A. J. & STRINGER, C. (2001). Human origins and ancient human DNA. *Science*, 292, 1655-1656.

COPPENS, Y. (1991), *L'évolution des hominidés, de leur locomotion et de leurs environnements*. In: COPPENS, Y. & SENUT, B. (Eds.), *Origine(s) de la bipédie chez les Hominidae, Cah. Paléoanthrop.*, CNRS, Paris (1991), pp. 295–301.

COQUEUGNIOT, H., HUBLIN, J.-J., VEILLON, F., HOUET, F. & JACOB, T. (2004), Early brain growth in *Homo erectus* and implications for cognitive ability, *Nature* 431:299-302

CORBALLIS, M. C. (2004), The origins of modernity: was autonomous speech the critical factor?, *Psychological Review* 111:543-552

CRACRAFT, J. (1989), *Speciation and its ontogeny: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation*. In OTTE D. & ENDLER J.A. (Eds.), *Speciation and its consequences*. Sunderland MA: Sinauer, pp 28–59.

CRESSIE, N. A. C. (1991), *Statistics for Spatial Data*, New York, USA: John Wiley & Sons

CROFT, W. (1990), *Typology and universals*, Cambridge : Cambridge University Press

CROW, T. (2002a), *Introduction*, in CROW (2002):1-20

CROW, T. (2002b), *Sexual selection, timing and an X-Y homologous gene: Did Homo*

sapines speciate on the Y chromosome?, In CROW (2002):197-216

CROW, T. (2002c), Handedness, language lateralisation and anatomical asymmetry: relevance of *protocadherin XY* to hominid speciation and the aetiology of psychosis. *British Journal of Psychiatry* 181, 295 – 297.X

CROW, T. (Ed.) (2002), *The Speciation of Modern Homo Sapiens*, Oxford: Ofxord University Press (*for* The British Academy)

CRYSTAL, D. (1975). *The English tone of voice: Essays in intonation, prosody and paralanguage.* London: Edward Arnold.

CURNOW, T. J. (2001), *What language features can be 'borrowed'*?, In AIKHENVALD, A. Y. & DIXON, R. M. W. (Eds.), *Areal diffusion and genetic inheritance, Oxford University Press*: NY, pp. 412-436

CURRAT, M., EXCOFFIER, L., MADDISON, W., OTTO, S. P. , RAY, N., WHITLOCK, M. C. & YEAMAN, S. (2006), Comment on ''Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens'' and ''Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans', *Science* 313:172

CUTLER, A., DAHAN, D., AND VAN DONSELAAR, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech* 40: 141-201.

DARNTON, J. (1996) *Neanderthal*, Random House Inc

DARWIN, C. (1872), *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, 6th Edition, John Murray, Albemarle Street: London. Online: http://pages.britishlibrary.net/charles.darwin/texts/origin_6th/origin6th_14.html

DAVIES, N. (1997), *Europe: a history*, Pimlico: London

DAWKINS, R. (1982), *The extended phenotype*, Oxford University Press

DAWKINS, R. (1990a), *The blind watchmaker*, Penguin Books

DAWKINS, R. (1990b), *The selfish gene*, Oxford University Press

DAWKINS, R. (1997), *Climbing Mount Improbable*, Penguin Books

DAWKINS, R. (2004), *The ancestor's tale: a pilgrimage to the dawn of life* (paperback), London: Phoenix/Orion Books

DE CEUNINCK, I. D., SCHNEIDER, S., LANDANEY, A. & EXCOFFIER, L. (2000), Inferring the impact of linguistic boundaries on population differentiation: application to the Afro-Asiatic-Indo-European case, *European Journal of Human Genetics* 8:750-756

DE LUMLEY, H., LORDKIPANIDZE, D., FÉRAUD, G., GARCIA, T., PERRENOUD, C., FALGUÈRES, C., GAGNEPAIN, J., SAOS, T., VOINCHET, P. (2002), Datation par la méthode ^{40}Ar / ^{39}Ar de la couche de cendres volcaniques (couche VI) de Dmanissi (Géorgie) qui a livré des restes d'hominidés fossiles de 1,81 Ma, *Comptes Rendus Palevol* 1:181-189

DE VAUS, D. (2002), *Analyzing social science data*, London: SAGE Publications

DEFRIES, J. C., & FULKER, D. W. (1988). Multiple regression analysis of twin data: Etiology of deviant scores versus individual differences. *Acta Geneticae Medicae et Gemellologicae*, 37, 205–216.

DEMETER, F., MANNI, F. & COPPENS, Y. (2003), Late Upper Pleistocene human peopling of the Far East: multivariate analysis and geographic patterns of variation, *Human*

Palaeontology and Prehistory, C. R. Palevol 2:625–638

DENCH, A. (2001), *Descent and diffusion: the complexity of the Pilbara situation*, In AIKHENVALD, A. Y. & DIXON, R. M. W. (Eds.), *Areal diffusion and genetic inheritance*, Oxford University Press: NY, pp.105-133

DENNELL, R. (2003), Dispersal and colonisation, long and short chronologies: how continuous is the Early Pleistocene record for hominids outside East Africa?, *Journal of Human Evolution* 45:421-440

DENNELL, R. & ROEBROEKS, W. (2005), An Asian perspective on early human dispersal from Africa, *Nature* 438:1099-1104

DESIGN R PACKAGE: FRANK E HARRELL JR (2005). *Design: Design Package*. R package version 2.0-12. http://biostat.mc.vanderbilt.edu/s/Design, http://biostat.mc.vanderbilt.edu/rms

DESMOND, A. & MOORE, J. (1992), *Darwin*, Penguin: London

DI GIACOMO, F., LUCA, F., POPA, L. O., AKAR, N., ANAGNOU, N., BANYKO, J., BRDICKA, R., BARBUJANI, G., PAPOLA, F., CIAVARELLA, G., CUCCI, F., DI STASI, F., GAVRILA, L., KERIMOVA, M. G., KOVATCHEV, D., KOZLOV, A. I., LOUTRADIS, A. *ET AL.* (2004), Y chromosomal haplogroup J as a signature of post-neolithic colonization of Europe, *Human Genetics* 115:357-371

DIAMOND, J. (1997), The language steamrollers, *Nature* 389:544-546

DIAMOND, J. (1998), *Guns, Germs and Steel: A short history of everybody for the last 13,000 years*, Vintage: London

DIAMOND, J. (2002), Evolution, consequences and future of plant and animal domestication, *Nature* 418:700-707.

DIAMOND, J. & BELLWOOD, P. (2003), Farmers and their languages: the first expansions, *Science* 300:597-603

DIFFLOTH, G. (2005), *The contribution of linguistic palaeontology to the homeland of Austro-Asiatic*, In SAGART, BLENCH & SANCHEZ-MAZAS (2005), pp.:77-80

DIMMENDAAL, G. J. (2001), *Areal diffusion versus genetic inheritance: an African perspective*, In AIKHENVALD, A. Y. & DIXON, R. M. W. (Eds.), *Areal diffusion and genetic inheritance*, Oxford University Press: NY, pp.358-392

DIXON, R. M. W. (1997), *The Rise and Fall of Languages*, Cambridge University Press: New York

DIXON, R. W. W. (2001), *The Australian linguistic area*, In AIKHENVALD, A. Y. & DIXON, R. M. W. (Eds.), *Areal diffusion and genetic inheritance*, Oxford University Press: NY, pp.64-104

DOBROVOLSKY, M. & KATAMBA, F., (1997), *Phonetics: the sounds of language*, In O'GRADY, DOBROVOLSKY & KATAMBA (Eds.) (1997), pp.:18-67

DOBZHANSKY, T. (1944), On species and races of living and fossil man, *American Journal of Physical Anthropology* 2:251-265

DOBZHANSKY, T. (1955), A review of some fundamental concepts and problems of population genetics, *Cold Spring Harbor Symposium on Quantitative Biology* 20:1-15

DOBZHANSKY, T. (1963) Possibility that *Homo Sapiens* Evolved Independently 5 Times Is Vanishingly Small, *Current Anthropology* 4:360, 364-367

DOLGOPOLSKY, A. (1999), *The Nostratic macrofamily: a short introduction*, In RENFREW, C. & NETTLE, D. (Eds.) (1999), pp.: 19-44

DONALD, M. (1999), *Les origines de l'esprit moderne: Trois étapes dans l'évolution de la culture et de la cognition*, DeBoeck Université

DONNELLY, M. J., PINTO, J., GIROD, R., BESANSKY, N. J. & LEHMANN, T. (2004), Revisiting the role of introgression vs shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the *Anopheles gambiae* complex, *Heredity* 92:61-68.

DOWLING, T. E. & SECOR, C. L. (1997), The role of hybridization and introgression in the diversification of animals, *Annual Review of Ecology and Systematics* 28:593-619.

DOWMAN, M., KIRBY, S. & GRIFFITH, T.L. (2006), *Innateness and culture in the evolution of language*, In CANGELOSI, SMITH & SMITH (Eds.) (2006).

DUARTE, C., MAURÍCIO, J., PETTITT, P. B., SOUTO, P., TRINKAUS, E., VAN DER PLICHT, H. & ZILHÃO, J. (1999), The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia, *Proceedings of the National Academy of Sciences of the USA* 96:7604-7609

DUMITRAȘCU, N. (2005), *Tehnicile proiective în evaluarea personalității*, Editura Trei, București

DUNBAR, R. (1996), *Grooming, gossip and the evolution of language*, London: Faber & Faber

DYEN, I., KRUSKAL, J. B. & BLACK, P. (1992), *An Indoeuropean classification: a lexicostatistical experiment*, Transactions of the American Philosophical Society, 82, Part 5

ECKHARDT, R. B., WOLPOFF, M. H. & THORNE, A. G. (1993), Multiregional Evolution, *Science* 262:973-974

EDGINGTON, E. S. (1987), *Randomization Tests* (2nd Edition), Marcel Dekker Inc.: NY

EDWARDS, A. W. F. (2003), Human genetic diversity: Lewontin's fallacy, *BioEssays* 25:798-801

EHRET, C. (1999), *Nostratic – or proto-human*?, In RENFREW, C. & NETTLE, D. (Eds.), pp.:93-112

ELDREDGE, N. & GOULD, S. J. (1972) *Punctuated equilibria: An alternative to phyletic gradualism*. In: SCHOPF., T. J. M. (Ed.), *Models in palaeobiology*, Freeman & Cooper: San Francisco.

ELIADE, M. (1981), *A History of Religious Ideas: From the Stone Age to the Eleusinian Mysteries* (Vol. 1), University of Chicago Press

ELLISON, M. T. & KIRBY, S. (2006), Measuring language divergence by intra-lexical comparison, *Proceedings of COLING-ACL*, 2006.

EMBLETON, S. (2000), *Lexicostatistics/Glottochronology: from Swadesh to Sankoff to Starostin to future horizons*, In RENFREW, MCMAHON & TRASK (2000), pp 143-165

ENARD, W., PRZEWORSKI, M., FISHER, S.E., LAI, C.S., WIEBE, V., KITANO, T., MONACO, A.P. & PÄÄBO, S. (2002), Molecular evolution of *FOXP2*, a gene involved in speech and language, *Nature* 418:869-872.

ESWARAN, V. (2002), A diffusion wave out of Africa: The mechanism of the modern human revolution?, *Current Anthropology* 43:749-773

EVANS, P. D., GILBERT, S. L., MEKEL-BOBROV, N., VALLENDER, E. J., ANDERSON, J. R., VAEZ-AZIZI, L. M., TISHKOFF, S. A., HUDSON, R. R. & LAHN, B. T. (2005), Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans, *Science* 309:1717 – 1720

EVANS, P.D., MEKEL-BOBROV, N., VALLENDER, E.J., HUDSON, R.R. & LAHN, B.T. (2006), Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from an archaic *Homo* lineage, *PNAS* Early edition, doi:10.1073.pnas.0606966103.

FAGAN, B. (2004), *The Long Summer: How Climate Changed Civilization*, Granta Books

FALK, D., HILDEBOLT, C., SMITH, K., MORWOOD, M. J., SUTIKNA, T., BROWN, P., JATMIKO, SAPTOMO, E. W., BRUNSDEN, B., PRIOR, F. (2005) The Brain of LB1, *Homo floresiensis Science* 308:242-245

FALK, D., HILDEBOLT, C., SMITH, K., MORWOOD, M. J., SUTIKNA, T., JATMIKO, SAPTOMO, E. W., BRUNSDEN, B. & PRIOR, F. (2006), Response to Comment on "The brain of LB1, *Homo floresiensis*", *Science* 312:999c.

FELSENFELD, S. (2002), Finding susceptibility genes for developmental disorders of speech: the long and winding road, *Journal of Communication Disorders* 35:329-345

FELSENSTEIN, J. (2005), *PHYLIP (Phylogeny Inference Package) version 3.6*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

FERRARO, G. (2001), *Cultural Anthropology – An Applied Perspective*, Wadsworth: Belmont, CA, USA.

FINLAYSON, C. (2005), Biogeography and evolution of the genus *Homo*, *Trends in Ecology and Evolution*, 20:457-463

FISHER, S. E., LAI, C.S. L. & MONACO, A. P. (2003), Deciphering the genetic basis of speech and language disorders, *Annual Review of Neuroscience* 26:57-80

FISHER, S. E., VARGHA-KHADEM, K. E., MONACO, A. P. & PEMBERY, M. E. (1998), Localization of a gene implicated in a severe speech and language disorder, *Nature Genetics* 18:168-170

FITCH, W. T. (2000) The evolution of speech: a comparative review, *Trends in Cognitive Sciences* 4:258-267

FORTIN, M.-J. & DALE, M. (2005), *Spatial analysis: A guide for ecologists*, Cambridge University Press: Cambridge, UK

FORTSON, B. W. (IV) (2004), *Indo-European language and Culture: an introduction*, Blackwell Publishing: UK

FRY, D. B. (1979) *The Physics of Speech.* Cambridge: Cambridge University Press.

FULLER, D. (2002), *An agricultural persepctive on Dravidian historical linguistics: archaeological crop packages, livestock and Dravidian crop vocabulary*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:191-213

GABORA, L. (2003), Cultural Focus: A cognitive explanation for the cultural transition of the Middle/Upper Paleolithic, *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Lawrence Erlbaum Associates.

GABOUNIA, L., DE LUMLEY, M.-A., VEKUA, A., LORDKIPANIDZE, D., DE LUMLEY, H. (2002), Découverte d'un nouvel hominidé à Dmanissi (Transcaucasie, Géorgie), *Comptes Rendus Palevol* 1:243-253

GABUNIA, L., VEKUA, A., LORDKIPANIDZE, D., SWISHER III, C. C., FERRING, R., JUSTUS, A., NIORADZE, M., TVALCHRELIDZE, M., ANTÓN, S. C., BOSINSKI, G., JÖRIS, O., DE LUMLEY, M.-A., MAJSURADZE, G., MOUSKHELISHVILI, A. (2000), Earliest Pleistocene Hominid Cranial Remains from Dmanisi, Republic of Georgia: Taxonomy, Geological Setting, and Age, *Science* 288:1019-1025

GABUNIA, L.K. & ABESALOM, K. V. (1995) , A Plio-Pleistocene hominid from Dmanisi, East Georgia, Caucasus. *Nature* 373:509–512

GARRIGAN, D., MOBASHER, Z., KINGAN, S. B., WILDER, J. A. & HAMMER, M. F. (2005a) Deep Haplotype Divergence and Long-Range Linkage Disequilibrium at Xp21.1 Provide Evidence That Humans Descend From a Structured Ancestral Population, *Genetics* 170:1849-1856

GARRIGAN, D., MOBASHER, Z., SEVERSON, T., WILDER, J. A. & HAMMER, M. F. (2005b), Evidence for archaic Asian ancestry on the human X chromosome, *Molecular Biology and Evolution* 22:189-192

GEARY, R. C. (1954), The contiguity ratio and statistical mapping. *The Incorporated Statistician* 5:115–145.

GERHART, J. & KIRSCHNER, M. (1997), *Cells, embryos and evolution: toward a cellular and developmental understanding of phenotypic and evolutionary adaptability*, Massachusetts: Blackwell Science

GIL, D. (1994), *Expressive Power*, In Asher, R. E. (Ed.), *The Encyclopaedia of Language and Linguistics*, Vol. 3, Pergamon Press: Oxford, pp. 1195-1198.

GIL, D. (2004), *Cross-linguistic variation in overall expressive power*, International Symposium on Typology of the Argument Structure and Grammatical Relations in Languages Spoken in Europe and North and Central Asia (LENCA-2), Kazan State University, May 11-14, 2004. Online: http://www.ling.helsinki.fi/uhlcs/LENCA/LENCA-2/information/datei/11-gil.pdf (November 2006).

GILBERT, S. F. (2000), *Developmental biology* (6th edition), Sunderland, Mass. : Sinauer

GILBERT, S. L., DOBYNS, W. B. & LAHN, B. T. (2005), Genetic links between brain development and brain evolution, *Nature Reviews: Genetics*, 6:581-590

GILBERT, W. H., WHITE, T. D. & ASFAW, B. (2003) *Homo erectus*, *Homo ergaster*, *Homo "cepranensis"* and the Daka cranium, *Journal of Human Evolution*, 45:255-259

GILLESPIE, J.H. (1984), The molecular clock my be episodic, *Proceedings of the National Academy of Sciences of the USA* 81: 8009-8013

GILLESPIE, N. A. & MARTIN, N. G. (2005), *Multivariate genetic analysis*. In EVERITT, B. S. & HOWELL, D. C. (Eds), *Encyclopedia of Statistics in Behavioral Science*: 1363-1370, Wiley: Chichester.

GOPNIK, M. & CRAGO, M. B. (1991), Familial aggregation of a developmental language disorder, *Cognition* 39:1-50.

GORDON, R. G., (Ed.) (2005) *Ethnologue: Languages of the World*, 15th edition. Dallas,

Tex.: SIL International. Online version: http://www.ethnologue.com/

Goren-Inbar, N., Alperson, N., Kislev, M. E., Simchoni, O., Melamed, Y., Ben-Nun, A., Werker, E. (2004) Evidence of Hominin Control of Fire at Gesher Benot Ya'aqov, Israel, *Science* 304:725-727

Gould, S. J. (1987), Nonoverlapping Magisteria, *Natural History* 96:16-22

Gould, S. J. (1988), Honorable men and women, *Natural History* 97:20-28

Gould, S. J. (2002), *The structure of evolutionary theory*, Cambridge MA:Harvard University Press

Greenberg, J. (1954), Studies in African linguistic classification VIII. Further remarks on method; revisions and corrections. *Southwestern Journal of Anthropology* 10:405-415

Greenberg, J. (1963a), *Languages of Africa*, Indiana University press: Bloomington

Greenberg, J. (1963b), *Some universals of grammar with particular reference to the order of meaningful elements*, In Greenberg, J. (Ed.), *Universals of languages*. MIT Press: Cambridge, Mass.

Greenberg, J. (1971), *The Indo-Pacific hypothesis*. In: Sebeok, T.A. (Ed.), *Current trends in linguistics, Vol. 8: Linguistics in Oceania*. Mouton: The Hague and Paris, pp. 807–871

Greenberg, J. (1987), *Language in the Americas*, Stanford University Press: Stanford

Greenberg, J. (1998), *The convergence of Eurasiatic and Nostratic*, In Salmons, J. C. & Joseph, B. D. (Eds.) (1998), *Nostratic: sifting the evidence*, John Benjamins Publishing Co.: Amsterdam, pp. 51-60

Greenberg, J. (2000), *Indo-European and Its Closest Relatives: The Eurasiatic Language Family*, Vol. I, *Grammar*. Stanford University Press: Stanford

Greenberg, J. (2002), *Indo-European and Its Closest Relatives: The Eurasiatic Language Family*, Vol. II, *Lexicon*. Stanford University Press: Stanford

Gregory, W. K. (1949), Franz Weidenriech, 1873-1948, *American Anthropologist* 51:85-90

Griffiths, T. & Kalish, M. (2005). A Bayesian view of language evolution by iterated learning. *In Proceedings of the 27th Annual Conference of the Cognitive Science Society.*

Gross, P. R. & Levitt, N. (1998), *Higher superstition: the academic left and its quarrels with science*, Baltimore: John Hopkins University Press

Groves, C.P. (2001), *Primate Taxonomy*. Smithsonian Institution Press: Washington, D.C

Grønbech, K. & Krueger, J. R. (1993), *An Introduction to Classical (Literary) Mongolian*, 3rd edition. Wiesbaden: Harrassowitz Verlag.

Grün, R., Stringer, C., McDermott, F., Nathan, R., Porat, N., Robertson, S., Taylor, L., Mortimer, G., Eggins, S. & McCulloch, M. (2005), U-series and ESR analyses of bones and teeth relating to the human burials from Skhul, *Journal of Human Evolution* 49:316-334

Güterbok, H. G. & Hoffner, H. A. (1997), *The Hittite Dictionary of the Oriental Institute of the University of Chicago, Volume P*, The Oriental Institute of the University of Chicago: Chicago (online http://oi.uchicago.edu/OI/DEPT/PUB/SRC/CHD/CHDP.pdf, June, 2006)

GUTHRIE, M. (1948), *The Classification of the Bantu Languages*. Oxford University Press

GUTHRIE, M. (1953), *The Bantu Languages of Western Equatorial Africa*. Oxford University Press

GUTIÉRREZ, G., SÁNCHEZ, D. & MARÍN, A. (2002). A reanalysis of the ancient mitochondrial DNA sequences recovered from Neanderthal bones. *Molecular Biology and Evolution*, 19, 1359-1366.

HAESLER, S., WADA, K., NSHDEJAN, A., MORRISEY, E.E., LINTS, T., JARVIS, E.D. & SCHARFF, C. (2004), *FoxP2* expression in avian vocal learners and non-learners, *Journal of Neuroscience* 24:3164-3175.

HAILE-SELASSIE, Y. (2001), Late Miocene hominids from the Middle Awash, Ethiopia, *Nature* 412, 178-181

HALLIBURTON, R. (2004), *Introduction to population genetics*, Upper Saddle River: Pearson Eduction Inc. (International Edition)

HALMOS, P. R. (2001) *Naive Set Theory*, New York: Springer-Verlag Inc.

HAMILTON, W. D. (1964), The genetical evolution of social behaviour. I, *Journal of Theoretical Biology* 7:1-52

HAMP, E. P. (1998), *Some draft principles for classification*, In SALMONS, J. C. & JOSEPH, B. D. (Eds.) (1998), *Nostratic: sifting the evidence*, John Benjamins Publishing Co.: Amsterdam, pp. 13-15

HANIHARA, T. (1996), Comparison of craniofacial features of major human groups, *American Journal of Physical Anthropology* 99:389-412

HARDING, R.M. & MCVEAN, G. (2004), A structured ancestral population for the evolution of modern humans, *Current Opinion in Genetics & Development*, 14:667-674

HARPENDING, H. & ESWARAN, V. (2005), Tracing modern human origins, *Science* 309:1995.

HARRISON, R. G. (1998) *Linking Evolutionary Pattern and Process – The relevance of species concepts for the study of speciation*, in HOWARD & BERLOCHER (Eds), 1998:19-31

HARVATI, K., FROST, S. R. & MCNULTY, K. P. (2004). Neanderthal taxonomy reconsidered: implications of 3D primate models of intra- and interspecific differences. *PNAS*, 101, 1147-1152.

HASPELMATH, M., DRYER, M. S., GIL, D. & COMRIE, B. (2005), *The World Atlas of Language Structures*, Oxford University Press

HASSAN, F. A. (2002), *Archaeology and linguistic diversity in North Africa*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:127-133

HAUSER, M.D., CHOMSKY, N. & FITCH, W.T. (2002), The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?, *Science* 298:1569-1579

HAVILAND, W. A. (2000), *Human Evolution and Prehistory*, 5th Edition, Harcourt College Publishers

HAWKS, J. & WOLPOFF, M. H. (2001), The four faces of Eve: hypothesis compatibility and human origins, *Quaternary International* 75:41-50

HAWKS, J. & WOLPOFF, M. H. (2003), Sixty years of modern human origins in the American Anthropological Association, *American Anthropologist* 105:87-98

HDGP: THE HUMAN DIVERSITY PANEL GENOTYPES (data used by ROSENBERG *ET AL.*, (2002)) and available online Dec. 2005 at http://research.marshfieldclinic.org/genetics/Freq/FreqInfo.htm and currently (an updated version) at http://rosenberglab.bioinformatics.med.umich.edu/diversity.html.

HENNEBERG, M. (2003) *Comment to HOLLIDAY (2003)*.

HENRY, D. O. (Ed.) (2003) *Neanderthals in the Levant: behavioral Organization and the beginnings of human modernity*, NY: Continuum

HENSHILWOOD, C. S. & MAREAN, C. W. (2003), The origin of modern human behavior – Critique of the models and their test implications, *Current Anthropology* 44:627-651

HERTZOG, P. J. & KOLA, I. (2001), *Overview: Gene knockouts*, in TYMMS, M. J. & KOLA, I. (Eds.), *Gene knockout protocols* (Methods in molecular biology: 158), Humana Press Inc.

HEY, J. (2001a) The mind of the species problem, *Trends in Ecology and Evolution* 16:326-329

HEY, J. (2001b), *Genes, Categories and Species – The evolutionary and cognitive causes of the species problem*, Oxford University Press

HIGHAM, C. (2002), *Languages and farming dispersals: Austroasiatic languages and rice cultivation*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:223-232

HIRST, D. & DI CRISTO, A. (1998), *A survey of intonation systems*. In HIRST, D. & DI CRISTO, A. (Eds) (1998), *Intonation Systems : A Survey of Twenty Languages*, Cambridge University Press: Cambridge, pp. 1-44

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800-803.

HOCKETT, B. & HAWS, J.A. (2005), Nutritional ecology and the human demography of Neandertal extinction, *Quaternary International* 137:21-34

HOLLIDAY, T. W. (2003), Species Concepts, Reticulation, and Human Evolution, *Current Anthropology* 44:653-673

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65-70.

HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383-386.

HORAN, D. R., BULTE, E. & SHOGREN, J. F. (2005), How trade saved humanity from biological exclusion: an economic theory of Neaderthal extinction, *Journal of Economic Behaviour & Organization* 58:1-29

HOURANI, A. (2002), *A history of the Arab peoples*, Faber and Faber Ltd: London

HOWARD, D. J. & BERLOCHER, S. H. (Eds) (1998), *Endless Forms – Species and speciation*, New York: Ofxord University Press

HOWELL, N., ELSON, J. L., TURNBULL, D. M. & HERRNSTADT, C. (2004), African Haplogroup L *mtDNA* Sequences Show Violations of Clock-like Evolution, *Molecular Biology and Evolution* 21:1843-1854

HOWITT, D. & CRAMER, D. (2003), *An introduction to statistics in psychology* (Revised 2nd edition), Essex: Pearson Education Ltd: Essex, England.

HROZNY, B. (1915), Die Losung des Hethetischen problem, *Mitteilungan der Deutschen Orient-Gesellschaft* 56:17-50

HUNT, K. D. (2003), The Single Species Hypothesis: Truly Dead and Pushing Up Bushes, or Still Twitching and Ripe for Resuscitation?, *Human Biology* 75:485-502

HURFORD, J. (2002), *Expression/induction models of language evolution: Dimensions and issues*, In BRISCOE, T. (Ed.), *Linguistic Evolution Through Language Acquisition*, Cambridge University Press

HURFORD, J. (2003), *The language mosaic and its evolution*, In Christiansen, M. & Kirby, S. (Eds.), *Language Evolution*, Oxford University Press, pp. 38-57.

HURFORD, J. R., STUDDERT-KENNEDY, M. & KNIGHT C. (Eds.) (1998), *Approaches to the Evolution of Language - Social and Cognitive Bases*, Cambridge University Press

HURLES, M. (2002), *Can the Hypothesis of Language/Agriculture Co-dispersal be* Tested *with Archaeogenetics*?, In BELLWOOD & RENFREW (Eds.) (2002), pp.:299-309

HURST, G. D. D. & WERREN, J. H. (2001), The role of selfish genetic elements in eukaryotic evolution, *Nature Reviews Genetics* 2:597-606

HURST, J. A., BARAITSER, M., AUGER, E., GRAHAM, F. & NORELL, S. (1990), An extended family with a dominantly inherited speech disorder, *Developmental Medicine and Child Neurology* 32:352-355

INOUE, K. & LUPSKI, J. R. (2003), Genetics and genomics of behavioral and psychiatric disorders, *Current Opinion in Genetics & Development* 13:303-309

IVĂNESCU, G. (2000), *Istoria Limbii Române*, Ed. Junimea, Iaşis

JACKSON, J. P. JR. (2001) "In Ways Unacademical": The Reception of Carleton S. Coon's The Origin of Races, *Journal of the History of Biology* 34:247–285

JACOB, T., INDRIATI, E., SOEJONO, R. P., HSÜ, K., FRAYER, D. W., ECKHARDT, R. B., KUPEREVAGE, A. J., THORNE, A. & HENNEBERG, M. (2006), Pygmoid Austromelanesian *Homo sapiens* skeletal remains from Liang Bua, Flores: population affinities and pathological abnormalities, *PNAS* 103:13421-13426.

JAEGER, J.-J. & MARIVAUX, L., (2005) Shaking the Earliest Branches of Anthropoid Primate Evolution, *Science* 310:244-245

JAEGER, J.-J., SOE, U. A. N., AUNG, U. A. K., BENAMMI, M., CHAIMANEE, Y., DUCROCQ, R.-M., TUNE, COL.T. , THEINF, U. T. & DUCROCQ, S. (1998), New Myanmar middle Eocene anthropoids: An Asian origin for catarrhines?, *Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie* 321:953-959

JAKOBSON, R. (1971), *Why "Mama" and "Papa"*? In BAR-ADON, A. & LEOPOLD, F. (Eds.), *Child language: A book of readings*, Prentice-Hall: Englewood Cliffs, NJ.

JENSEN, A. R. (1998), *The g factor: The science of mental ability*, Greenwood Press

JOBLING, M.A., TYLER-SMITH, C. & HURLES, M. (2004), *Human Evolutionary Genetics: Origins, Peoples and Disease*, Garland Science: NY

JOLLY, C. J. (2001), A Proper Study for Mankind: Analogies From the Papionin Monkeys and Their Implications for Human Evolution, *Yearbook OF Physical Anthropology* 44:177–204

JOSEPH, B.D. (1999), *Romanian and the Balkans: Some Comparative Perspectives*. In

EMBLETON, S., JOSEPH, J. & NIEDEREHE, H.-J. (Eds.) *The Emergence of the Modern Language Sciences. Studies on the Transition from Historical-Comparative to Structural Linguistics in Honour of E.F.K. Koerner*. Volume 2: Methodological Perspectives and Applications. John Benjamins: Amsterdam, pp. 218-235

KIDDER, J. H. & DURBAND, A. C. (2004) A re-evaluation of the metric diversity within *Homo erectus*, *Journal of Human Evolution*, 46:299-315

KIMURA, M. (1968), Evolutionary rate at the molecular level, *Nature* 217:624-626.

KIMURA, M. (1983), *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KIRBY, S. (1999), *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford University Press.

KIRBY, S. (2000), *Syntax without Natural Selection: How compositionality emerges from vocabulary in a population of learners*, In KNIGHT, STUDDERT-KENNEDY & HURFORD (Eds.), pp. 303-323

KIRBY, S. (2001), Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* 5:102-110

KIRBY, S. & HURFORD, J. (2001), *The Emergence of Linguistic Structure: an Overview of the Iterated Learning Model*, In PARISI, D. & CANGELOSI, A. (Eds.), *Simulating the Evolution of Language*, Berlin: Springer Verlag

KLEIN, R. G. (1999), *The Human Career: Human Biological and Cultural Origins*, 2nd Edition. Chicago: University of Chicago Press

KNIGHT, C., STUDDERT-KENNEDY, M. & HURFORD, J.R. (Eds.) (2000), *The Evolutionary Emergence of Language – Social function and the origins of linguistic form*, Cambridge University Press

KOLATCH, E., TOYE, J. & DORR, B. (2004), *Look Alike/ Sound Alike: Algorithms for Assessing Drug Name Similarities*, Project Performance Corporation: White papers, http://www.ppc.com/modules/knowledgecenter/lasa.pdf

KOLLER, C. (2005), White vs. Black: Wissenschaftliche Argumente für die Rassentrennung in den USA, Hausarbeit im Rahmen des Seminars «Die Wissenschaft von der Rasse», *Historisches Seminar, Universität Zürich*, WS 04/05

KONDRAK, G. (2002); *Algorithms for Language Reconstruction*, PhD Thesis, University of Toronto, http://www.cs.ualberta.ca/~kondrak/papers/thesis.pdf

KOPTJEVSKAJA-TAMM, M. (2006), *The Circle That Won't Come Full: Two Potential Isoglosses in the Circum-Baltic Area*, In MATRAS, Y., MCMAHON, A. & VINCENT, N. (Eds.) (2006), *Convergence in Historical and Typological Perspective*, Palgrave Macmillan, pp. 182-226. Online http://www.ling.su.se/staff/tamm/aff_CB_Matras_fin.pdf, August 2006 (page numbers refer to this online version)

KRAMER, A., CRUMMETT, T. L. & WOLPOFF, M. H. (2001), Out of Africa and into the Levant: replacement or admixture in Western Asia?, *Quaternary International* 75:51-63

KRAYTSBERG,Y., SCHWARTZ, M., BROWN, T.A., EBRALIDSE, K., KUNZ, W.S., CLAYTON, D.A., VISSING, J. & KHRAPKO, K. (2004), Recombination of Human Mitochondrial DNA, *Science* 304:981

KRINGS, M., CAPELLI, C., TSCHENTCHER, F., GEISERT, H., MEYER, S., VON HAESELER, A., GROSSSCHMIDT, K., POSSNERT, G., PAUNOVIC, M. & PÄÄBO, S. (2000). A view of Neandertal genetic diversity. *Nature Genetics*, 26, 144-146.

KRINGS, M., GEISERT, H., SCHMITZ, R. W., KRAINITZKI, H. & PÄÄBO, S. (1999), DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen, *Proceedings of the National Academy of Sciences of the USA* 96:5581-5585

KRINGS, M., STONE, A., SCHMITZ, R. W., KRAINITZKI, H., STONEKING. M. & PÄÄBO, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell*, 90, 19-30.

KRISTOFFERSEN, G. (2003), *The tone bearing unit in Swedish and Norwegian tonology*. In JACOBSEN, H.G., BLESES, D., MADSEN, T.O. & THOMSEN, P. (Eds.) (2003): *Take Danish - for instance. Linguistic studies in honour of Hans Basbøll*. University Press of Southern Denmark: Odense, pp. 189-198. Online http://nora.hd.uib.no/NTT/tbu-paper.pdf, August 2006

KRISTOFFERSEN, G. (2006), *Dialect variation in East Norwegian Tone*. In GUSSENHOVEN, C. & RIAD, T. (Eds.) (2006), *Tones and Tunes: Studies in Word and Sentence Prosody*. Mouton de Gruyter: Berlin. Online http://nora.hd.uib.no/NTT/EnDialectvariation.pdf, August 2006

KUTSCHERA, U. & NIKLAS, K.J. (2005), Endosymbiosis, cell evolution, and speciation, *Theory in Biosciences* 15:1-24

LAI, C. S. L., GERRELLI, D., MONACO, A. P., FISHER, S. E. & COPP, A. J. (2003), *FOXP2* expression during brain development coincides with adult sites of pathology in a severe speech and language disorder, *Brain* 126:2455-2462

LAI, C. S. L.; FISHER, S. E.; HURST, J. A.; LEVY, E. R.; HODGSON, S.; FOX, M.; JEREMIAH, S.; POVEY, S.; JAMISON, D. C.; GREEN, E. D.; VARGHA-KHADEM, F.; MONACO, A. P. (2000), The *SPCH1* region on human 7q31: genomic characterization of the critical interval and localization of translocations associated with speech and language disorder. *American Journal of Human Genetics* 67:357-368.

LAI, C. S., FISHER, S. E., HURST, J. A., VARGHA-KHADEM, F. & MONACO, P. (2001), A forkhead-domain gene is mutated in a severe speech and language disorder, *Nature* 413:519-523

LALUEZA-FOX, C., LOURDES SAMPIETRO, M., CARAMELLI, D., PUDER, Y., LARI. M., CALAFELL, F., MARTÍNEZ-MAZA, C., BASTIR, M., FORTEA, J., DE LA RASILLIA, M, BERTRANPETIT, J & ROSAS, A. (2005). Neandertal evolutionary genetics: mitochondrial DNA data from the Iberian peninsula. *Molecular Biology and Evolution*, 22, 1077-1081.

LANYON, S. J. (2006), *A saltationist approach for the evolution of human cognition and language*, in CANGELOSI, A., SMITH, A. D. M., SMITH, K. (Eds.), *The Evolution of Language: Proceedings of the 6th International Conference (EVOLANG6), Rome, Italy*. London: World Scientific, pp. 176-183.

LASS, R. (1997), *Historical linguistics and language chang*e, Cambridge University Press

LAZARD, G. (1992), *A Grammar of Contemporary Persian*. Mazda Publishers in association with Bibliotheca Persica, Costa Mesa CA.

LE MAY, M. (1975), The language capability of Neanderthal man, *American Journal of Physical Anthropology* 42:9-14

LEAKY, L. (1935), *The Stone Age Races of Kenya*. London: Oxford University Press

LEDER, M., HEFFRON, J. & THE EDITORS OF 'WRITER'S DIGEST' (Eds.) (2002), *The Complete Handbook of Novel Writing*, Writer's Digest Books

LEE, S.-H. & WOLPOFF, M. H. (2003), The pattern of evolution in Pleistocene human brain size, *Paleobiology* 29:186-196

LEGENDRE, P. AND LEGENDRE, L. (1998) *Numerical Ecology*. 2nd English Edition. Elsevier.

LEIGH, S. R. (2006), Brain ontogeny and life history in *Homo erectus*, *Journal of Human Evolution*, 50:104-108

LEVENSHTEIN, V. I. (1965); Binary codes capable of correcting deletions, insertions and reversals, *Doklady Akademii Nauk SSSR* 163:845-848

LEWIN, B. (2004), *Genes VIII*, Upper Saddle River: Pearson Prentice Hall

LEWIN, R. (1988), *Principles of human evolution – a core textbook*, Blackwell Science

LEWONTIN, R. C. (1972), *The apportionment of human diversity*, in DOBZHANSKY T., HECHT M. K. & STEERE, W. C. (Eds.), *Evolutionary biology 6*, NY:Appleton-Century-Crofts:381-398

LI, H., YAMAGATA, T., MORI, M. & MOMOI, M.Y. (2005), Absence of causative mutations and presence of autism-related allele in FOXP2 in Japanese autistic patients, *Brain and Development* 27:207-210.

LIÉGEOIS, F., BALDEWEG, T., CONNELLY, A., GADIAN, D. G., MISHKIN, M. & VARGHA-KHADEM, F. (2003), Language fMRI abnormalities associated with *FOXP2* gene mutation, *Nature Neuroscience* 6:1230-1237

LOEWENTHAL, K.M. (2001), *An introduction to psychological tests and scales*, Psychology Press Lt.: East Sussex, UK

LOHR, M. (2000), *New approaches to lexicostatistics and glottochronology*, In RENFREW, MCMAHON & TRASK (2000), pp 209-222

LONG, J. C. & KITTLES, R. A. (2003), Human Genetic Diversity and the Nonexistence of Biological Races, *Human Biology* 75:449-471

LOTSY, J. P. (1925) Species or linneon, *Genetics* 7:487-506

MACDERMOT, K.D., BONORA, E., SYKES, N., COUPE, A.M., LAI, C.S., VERNES, S.C., VARGHA-KHADEM, F., MCKENZIE, F., SMITH, R.L., MONACO, A.P. & FISHER, S.E. (2005), Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits, *American Journal of Human Genetics* 76:1074-1080.

MACEACHERN, S. (2000), Genes, Tribes, and African History, *Current Anthropology* 41:357-384.

MADDIESON, I. (2005), *Tone*, in HASPELMATH, M., DRYER, M. S., GIL, D. & COMRIE, B. (2005), *The World Atlas of Language Structures*, Oxford University Press

MADDISON, D.R. (1991), African origin of mitochondrial DNA re-examined. *Systematic Zoology*, 40:355-363.

MALLEGNI, F., CARNIERI, E., BISCONTI, M., TARTARELLI, G., RICCI, S., BIDDITTU, I., SEGRE, A. (2003), *Homo cepranensis sp. nov.* and the evolution of African-European Middle Pleistocene hominids, *Human Palaeontology and Prehistory / Paléontologie humaine et préhistoire* 2:153-159

MALLET, J. (2005), Hybridization as an invasion of the genome, *TRENDS in Ecology and*

Evolution 20:229-237.

MALLORY, J. P. (1991), *In Search of the Indo-Europeans: Language, Archaeology and Myth*, Thames and Hudson: London

MANTEL, N. (1967) The detection of disease clustering and a generalized regression approach, *Cancer Research* 27:209-220.

MAPS OF WORLD, http://www.mapsofworld.com/lat_long/index.html

MAPS R PACKAGE: ORIGINAL S CODE BY RICHARD A. BECKER AND ALLAN R. WILKS. R VERSION BY RAY BROWNRIGG ENHANCEMENTS BY THOMAS P MINKA (2006). *maps: Draw Geographical Maps*. R package version 2.0-31.

MARCUS, G. F. & FISHER, S. E. (2003), *FOXP2* in focus: what can genes tell us about speech and language?, *TRENDS in Cognitive Sciences* 7:257-262

MARGULIS, L & SAGAN, D. (1997), *Microcosmos: Four Billion Years of Microbial Evolution*, University of California Press

MARIVAUX, L., ANTOINE, P.-O., BAQRI, S. R. H., BENAMMI, M., CHAIMANEE, Y., CROCHET, J.-Y., DE FRANCESCHI, D., IQBAL, N., JAEGER, J.-J., MÉTAIS, G., ROOHI, G. & WELCOMME, J.-L. (2005), Anthropoid primates from the Oligocene of Pakistan (Bugti Hills): Data on early anthropoid evolution and biogeography, *PNAS* 102:8436-8441

MARSHAK, S. (2005), *Earth: Portrait of a Planet*, W.W. Norton & Co. Ltd.

MARTIN, R. D., MACLARNON, A. M., PHILLIPS, J. L., DUSSUBIEUX, L., WILLIAMS, P. R. & DOBYNS, W. B. (2006), Comment on "The brain of LB1, *Homo floresiensis*", *Science* 312:999b.

MASICA,C. P. (1991), *The Indo-Aryan Languages*. Cambridge University Press.

MATISOFF, J. A. (1990), On Megalocomparison, *Language* 66:106-120

MATISOFF, J. A. (2000), *On the uselessness of glottochronology for the subgrouping of Tibeto-Burman*, In RENFREW, MCMAHON & TRASK (2000), pp 333-373

MAU, B.-L., LEE, H.-M. & TZEN, C.-Y. (2005) Identification of Human-Specific Adaptation Sites of *ATP6*, *Annals of the New York Academy of Sciences* 1042:142-147

MAYNARD-SMITH, J. & SZATHMÁRY, E. (1995), *The major transitions in evolution*, New York: Oxford University Press

MAYR, E. (1942), *Systematics and the origin of species*, New York: Columbia University Press

MAYR, E. (1963), *Animal species and evolution*, Cambridge: Belknap Press

MCBREARTY, S. & BROOKS, A. S. (2000), The revolution that wasn't: a new interpretation of the origins of modern human behavior, *Journal of Human Evolution* 39:453-563

MCKENNA, M. C. & BELL, S. K. (1997), *Classification of Mammals Above the Species Level.* Columbia University Press, New York

MCMAHON, A. & MCMAHON, R. (2005), *Language classification by numbers*, Oxford University Press.

MCMAHON, R. (2004), Genes and languages, *Community Genetics* 7:2-13

MEKEL-BOBROV, N., EVANS, P. D., GILBERT, S. L., VALLENDER, E. J., HUDSON, R. R. & LAHN, B. T. (2006), Response to Comment on "Ongoing Adaptive Evolution of *ASPM*, a Brain Size Determinant in Homo sapiens" and "*Microcephalin*, a Gene Regulating Brain

Size, Continues to Evolve Adaptively in Humans'', *Science* 313:172b

MEKEL-BOBROV, N., GILBERT, S. L., EVANS, P. D., VALLENDER, E. J., ANDERSON, J. R., HUDSON, R. R., TISHKOFF, S. A. & LAHN, B. T. (2005), Ongoing Adaptive Evolution of *ASPM*, a Brain Size Determinant in Homo sapiens, *Science* 309:1720 – 1722

MELLARS, P. (2005), The impossible coincidence. A single-species model for the origins of modern human behavior in Europe, *Evolutionary Anthropology* 14:12-27

MELOY, J. R., ACKLIN M. W., GACONO, C. B., MURRAY, J. F. & PETERSON, C. A., (Eds.) (1997), *Contemporary Rorschach Interpretation*, Lawrence Erlbaum Ass. Inc.

MERESCHKOWSKY, C. (1905), Uber Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol. Centralbl.* 25 593–604, 689–691.

MILITAREV, A. (2002), *The prehistory of a dispersal: the Proto-Afrasian (Afroasiatic) farming lexicon*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:135-150

MISHMAR, D., RUIZ-PESINI, E., GOLIK, P., MACAULAY, V., CLARK, A. G., HOSSEINI, S., BRANDON, M., EASLEY, K., CHEN, E., BROWN, M. D., SUKERNIK, R. I., OLCKERS, A. & WALLACE, D. C. (2003), Natural selection shaped regional *mtDNA* variation in humans, *PNAS* 100:171-176

MITHEN, S. (1996), *The prehistory of the mind: A search for the origins of art, science and religion*. London: Thames & Hudson

MITHEN, S. (2003), *After the ice: a global human history*, Phoenix: Orion Books Ltd: UK

MOILANEN, J. S., FINNILA, S. & MAJAMAA, K. (2003), Lineage-Specific Selection in Human *mtDNA*: Lack of Polymorphisms in a Segment of *MTND5* Gene in Haplogroup J, *Molecular Biology and Evolution*, 20:2132-2142

MONTAGU, A. (1963) What Is Remarkable About Varieties Of Man Is Likeness, Not Differences, *Current Anthropology* 4:361-364

MORAN, P. (1948), The interpretation of statistical maps. *Journal of the Royal Statistical Society B* 10:243-289

MORAN, P. (1950), Notes on continuous stochastic phenomena. *Biometrika*, 37:17–23

MORTENSEN, D. (2004), *Preliminaries to Mong Leng (Hmong Njua) Phonology*. Unpublished, online http://ist-socrates.berkeley.edu/~dmort/mong_leng_phonology.pdf September 2006.

MORWOOD, M. J., BROWN, P., JATMIKO, SUTIKNA, T., SAPTOMO, E. W., WESTAWAY, K. E., DUE, R. A., ROBERTS, R. A., MAEDA, T., WASISTO, S. & DJUBIANTONO, S. (2005) Further evidence for small-bodied hominins from the Late Pleistocene of Flores, Indonesia, *Nature* 437:1012-1017

MORWOOD, M. J., O'SULLIVAN, P. B., AZIZ, F. & RAZA, A. (1998) Fission-track ages of stone tools and fossils on the east Indonesian island of Flores, *Nature* 392:173-176

MORWOOD, M. J., SOEJONO, R. P., ROBERTS, R. G., SUTIKNA, T., TURNEY, C. S. M., WESTAWAY, K. E., RINK, W. J., ZHAO, J.-X., VAN DEN BERGH, G. D., ROKUS AWE DUE, HOBBS, D. R., MOORE, M. W., BIRD, M. I. & FIFIELD, L. K. (2004), Archaeology and age of a new hominin from Flores in eastern Indonesia, *Nature* 431:1087-1091

MOURA, A. C. DE A. & LEE, P. C. (2004), Capuchin Stone Tool Use in Caatinga Dry Forest, *Science* 306:1909

MUGANE, J. M. (1997), *A Paradigmatic Grammar of Gikuyu*. Stanford: CLSI Publications.

NEI, M. (1972), Genetic Distance between Populations, *The American Naturalist* 106:283-292

NETTLE, D. (1998), Explaining global patterns of language diversity, *Journal of anthropological archaeology* 17:354-374

NETTLE, D. (1999a), *Linguistic diversity*, Oxford University Press

NETTLE, D. (1999b), Linguistic diversity of the Americas can be reconciled with a recent colonization, *PNAS* 96:3325-3329

NETTLE, D. (1999c), Is the rate of linguistic change constant?, *Lingua* 108:119-136

NETTLE, D. (2000), *Linguistic diversity, population spread and time depth*, In RENFREW, MCMAHON & TRASK (2000), pp 665-677

NETTLE, D. & HARRISS, L. (2003), Genetic and Linguistic Affinities between Human Populations in Eurasia and West Africa, *Human Biology* 75:331–344

NEWBURY, D. F., BISHOP, D. V. M. & MONACO, A. P. (2005), Genetic influences on language impairment and phonological short-term memory, *TRENDS in Cognitive Sciences* 9:528-534

NEWBURY, D.F., BONORA, E., LAMB, J.A., FISHER, S.E., LAI, C.S., BAIRD, G., JANNOUN, L., SLONIMS, V., STOTT, C.M., MERRICKS, M.J., BOLTON, P.F., BAILEY, A.J., MONACO, A.P. & INTERNATIONAL MOLECULAR GENETIC STUDY OF AUTISM CONSORTIUM (2002), *FOXP2* is not a major susceptibility gene for autism or specific language impairment, *American Journal of Human Genetics* 70:1318-1327.

NGRIP (2006), *Greenland Ice Core Chronologay 2005 (GICC05)* released 10 March 2006, available http://www.glaciology.gfy.ku.dk/data/GICC05_20y_10march2006.tx (http://www.glaciology.gfy.ku.dk/ngrip/index_eng.htm) April 2006

NICHOLS, J. (1992), *Linguistic Diversity in Space and Time*, University of Chicago Press: Chicago.

NICHOLS, J. (1997), Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26.359-84

NICHOLS, J. (2000), *Estimating dates of early American colonization events*, In RENFREW, MCMAHON & TRASK (2000), pp 643-663

NNET R PACKAGE: VENABLES, W. N. & RIPLEY, B. D. (2002) *Modern Applied Statistics with S.* Fourth Edition. Springer, New York. ISBN 0-387-95457-0

NORDBORG, M. (1998). On the probability of Neanderthal ancestry. *American Journal of Human Genetics*, 63, 1237-1240.

NORDBORG, M. (2004), *Were Neanderthals and anatomically modern humans different species*?, In JOBLING, HURLES & TYLER-SMITH (2004), Box 8.11:264

O'BRIEN E.K., ZHANG, X., NISHIMURA, C., TOMBLIN, J. B., & MURRAY, J. C. (2003), Association of specific language impairment (SLI) to the region of 7q31. *American Journal of Human Genetics* 72:1536-1543.

O'GRADY, W., DOBROVOLSKY, M. & KATAMBA, F. (Eds.) (1997), *Contemporary linguistics: An introduction*, Pearson Education: UK

O'SULLIVAN, P. B., MORWOOD, M., HOBBS, D., SUMINTO, F. A., SITUMORANG, M., RAZA, A. &

MAAS, R. (2001) Archaeological implications of the geology and chronology of the Soa basin, Flores, Indonesia, *Geology* 29:607-610

ODLING-SMEE, F.J., LALAND, K.N. & FELDMAN, M.W. (2003), *Niche construction – the neglected process in evolution*, Princeton University Press

OHTA, T. (1996), The current significance and standing of neutral and nearly neutral theories, *BioEssays* 18:673-677

OKABE, A., BOOTS, B. & SUGIHARA, K. (1992), *Spatial tessellation – Concepts and applications of Voronoi diagrams*, John Wiley & Sons Ltd.: West Sussex, England

OPPENHEIMER, S. (2004), *Out of Eden: The peopling of the world*, London: Robinson

ORGANISED PHONOLOGY DATA: *Nasioi [government spelling] (Naasioi [language spelling]) Language [NAS] Kieta – North Solomons Province* (online http://pnglanguages.org/png/LangResource/0000268/Nasioi.pdf)

OSIER M.V., CHEUNG K.H., KIDD J.R., PAKSTIS A.J., MILLER P.L. & KIDD K.K. (2002), ALFRED: an allele frequency database for Anthropology. *American Journal of Physical Anthropology* 119:77-83.

OSTLER, N. (2005), *Empires of the word: a Language History of the Worlds*, Harper Collins Publishers: London

OVCHINNIKOV, I. V., GÖTHERSTRÖM, A., ROMANOVA, G. P., KHARITONOV, V. M., LIDÉN. K. & GOODWIN, W. (2000). Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, 404, 490-493.

PANGER, M. A., BROOKS, A. S., RICHMOND, B. G. & WOOD, B. (2002), Older Than the Oldowan? Rethinking the Emergence of Hominin Tool Use, *Evolutionary Anthropology* 11:235-245

PARKER, A.R. (2006a), *Evolving the narrow language faculty: Was recursion the pivotal step*?, In CANGELOSI, SMITH & SMITH (Eds.), pp. 239-246

PARKER, A.R. (2006b), *Evolution as a Constraint on Theories of Syntax: The Case Against Minimalism*. PhD thesis, University of Edinburgh.

PATERSON, H.E.H. (1985) *The recognition concept of species*. In VRBA, E. S. (Ed.), *Species and speciation*, Pretoria: Transvaal Museum 21-29

PATTERSON, N., RICHTER, D. J., GNERRE, S., LANDER, E. S. & REICH, D. (2006), Genetic evidence for complex speciation of humans and chimpanzees, *Nature* 441:1103-1108.

PAYSEUR, B. A. & NACHMAN, M. W. (2005), The genomics of speciation: investigating the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus*, *Biological Journal of the Linnean Society* 84:523–534.

PAYSEUR, B. A., KRENZ, J. G. & NACHMAN, M. W. (2004), Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice, *Evolution* 58:2064–2078.

PEDERSEN, H. (1931), *The discovery of language: linguistic science in the nineteenth century* (cited in BOMHARD (1998)).

PENCHOEN, T. G. (1973), *Tamazight of the Ayt Ndhir*. Los Angeles: Undena Publications.

PENNISI, E. (2006), The dawn of stone age genomics, *Science* 314:1068-1071

PENROSE, R. (1989), *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*, Oxford University Press

PETTITT, P. B. (2002) The Neanderthal dead: exploring mortuary variability in Middle Palaeolithic Eurasia, *Before Farming* 1:1-19

PIAZZA, A., RENDINE, S., MINCH, E., MENOZZI, P., MOUNTAIN, J. & CAVALLI-SFORZA, L. L. (1995), Genetics and the origin of European languages, *PNAS* 92:5836-5840

PINKER, S. (1995), *The language instinct*, Penguin Books Ltd.

PINKER, S. (1997), *How the mind works*, Allen Lane The Penguin Press

PINKER, S. & JACKENDOFF, R. (2005), The faculty of language: what's special about it?, *Cognition* 95:201-236

PLOMIN, R. & KOVAS, Y. (2005), Generalist genes and learning disabilities, *Psychological Bulletin* 131:592-617

PLOMIN, R., COLLEDGE, E. & DALE, P. S. (2002), Genetics and the development of language disabilities and abilities, *Current Paediatrics* 12:419-424

PLOMIN, R., DEFRIES, J. C., MCCLEARN, G. E. & MCGUFFIN, P. (2001), *Behavioral Genetics* (4th edition), New York: Worth Publishers

PLOTRIX R PACKAGE: JIM LEMON, BEN BOLKER, SANDER OOM, EDUARDO KLEIN, BARRY ROWLINGSON AND HADLEY WICKHAM (2006). *plotrix: Various plotting functions*. R package version 2.1-2.

POLONI, E. S., SEMINO, O., PASSARINO, G., SANTACHIARA-BENERECETTI, A. S., DUPANLOUP, I., LANGANEY, A. & EXCOFFIER, L. (1997), Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics, *American Journal of Human Genetics* 61:1015-1035

POPPER, K. (2002), *The Logic of Scientific Discovery*, Routledge

POWLEDGE, T. M. (2005) Skullduggery – The discovery of an unusual human skeleton has broad implications, *European Molecular Biology Organization: EMBO reports: Science & Society* 6:609-612

R DEVELOPMENT CORE TEAM (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

RAAUM, R. L., STERNER, K. N., NOVIELLO, C. M., STEWART, C.-B. & DISOTELL, T. R. (2005), Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence, *Journal of Human Evolution* 48:237-257

RAJEEVAN H., OSIER M.V., CHEUNG K.H., DENG H., DRUSKIN L., HEINZEN R., KIDD J.R., STEIN S., PAKSTIS A.J., TOSCHES N.P., YEH C.C., MILLER P.L. & KIDD K.K. (2003), ALFRED – the ALlele FREquency Database – update. *Nucleic Acids Research* 31:270-271

RAMER, A. M., MICHALOVE, P. A., BAERTSCH, K. S. & ADAMS, K. L. (1998), *Exploring the Nostratic hypothesis*, In SALMONS & JOSEPH (Eds.), pp.:61-84

RANGANATH, H. A. & ARUNA, S. (2003), Hybridization, transgressive segregation and evolution of new genetic systems in *Drosophila*, *Journal of Genetics* 82:163-177.

REID, L. A. (2005), *The current status of Austric: A review and evaluation of the lexical and*

morhposyntactic evidence, In SAGART, BLENCH & SANCHEZ-MAZAS (2005), pp.:132-160

RELETHFORD, J. (2001), *Genetics and the search for modern human origins*, New York: Wiley-Liss

RELETHFORD, J. (2003), *Reflections of our past: How human history is revealed in our genes*, Cambridge, MA: Westview Press

RELIMP R PACKAGE: DAVID FIRTH (). *relimp: Relative Contribution of Effects in a Regression Model*. R package version 0.9-6. http://www.warwick.ac.uk/go/relimp, http://www.warwick.ac.uk/go/dfirth

RENFREW, C. (1991), Before Babel: speculations on the origins of linguistic diversity, *Cambridge Archaeological Journal* 1:3-23

RENFREW, C. (1999), *Nostratic as a linguistic macrofamily*, In RENFREW, C. & NETTLE, D. (Eds.), pp.:3-18

RENFREW, C. (2002), *'The Emerging Synthesis': the Archaeogenetics of Farming/Language Dispersals and other Spread Zone*s, In BELLWOOD & RENFREW (Eds.) (2002), pp.:3-16

RENFREW, C. & NETTLE, D. (1999), *Foreword*, In RENFREW, C. & NETTLE, D. (Eds.), p.:vii

RENFREW, C. & NETTLE, D. (Eds.) (1999), *Nostratic: examining a linguistic macrofamily*, The McDonald Institute for archaeological research, Oxbow Books: Oxford

RENFREW, C. MCMAHON, A. & TRASK, L. (Eds.) (2000), *Time depth in historical linguistics*, The McDonald Institute for Archaeological Research, Cambridge: Cambridge University Press

RIAD, T. (1996), Remarks on the Scandinavian tone accent typology, *Nordlyd* 24:129-156.

RIAD, T. (1998), The origin of Scandinavian tone accents, *Diachronica* 15:63-98

RINGE, D. (1998), *Probabilistic evidence for Indo-Uralic*, In SALMONS & JOSEPH (Eds.), pp.:153-39216

ROKAS, A., LADOUKAKIS, E. & ZOUROS, E. (2003), Animal mitochondrial DNA recombination revisited, *Trends in Ecology and Evolutio*n 18:411-417

RON, H. & LEVI, S. (2001), When did hominids first leave Africa?: New high-resolution magnetostratigraphy from the Erk-el-Ahmar Formation, Israel, *Geology* 29:887-890

ROSENBERG, N. A., PRITCHARD, J. K., WEBER, J. L., CANN, H. M., KIDD, K. K., ZHIVOTOVSKY, L. A. & FELDMAN, M. W. (2002), Genetic Structure of Human Populations, *Science* 298:2381-2385

ROSSER, Z. H., ZERJAL, T., HURLES, M. E., ADOJAAN, M., ALAVANTIC, D., AMORIM, A. *ET AL.*, (2000), Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language, *American Journal of Human Genetics* 67:1526-1543

ROUSSET, F. (2003). Effective size in simple metapopulation models. *Heredity*, 91, 107-111.

RUHLEN, M. (1991), *A Guide to the World's Languages*, Vol. I: *Classification* (with a postscript on recent developments). Arnold: London.

RUHLEN, M. (1994), *On the Origin of Languages: Studies in Linguistic Taxonomy*. Stanford, Stanford University Press

SAGART, L., BLENCH, R. & SANCHEZ-MAZAS, A. (2005), *The Peopling of East Asia: Putting*

together archaeology, linguistics and genetics, Routledge Curzon: NY

Salmons, J. C. & Joseph, B. D. (1998), *Introduction*, In Salmons, J. C. & Joseph, B. D. (Eds.) (1998), pp. 1-11

Salmons, J. C. & Joseph, B. D. (Eds.) (1998), *Nostratic: sifting the evidence*, John Benjamins Publishing Co.: Amsterdam

Saltzburger, W., Baric, S. & Sturmbauer, C. (2002), Speciation via introgressive hybridization in East African cichlids?, *Molecular Ecology* 11:619-625.

Sanderson, M.J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach, *Molecular Biology and Evolution* 19:101-109.

Sanjuan, J., Tolosa, A., Gonzalez, J.C., Aguilar, E.J., Molto, M.D., Najera, C. & de Frutos, R. (2005), *FOXP2* polymorphisms in patients with schizophrenia, *Schizophrenia Research* 73:253-256.

Sarich, V. & Miele, F. (2004), *Race: the reality of human differences*, Westview Press: Cambridge, USA

Sarich, V.M. and Wilson, A.C. (1967), Immunological time scale for hominid evolution. *Science* 158: 1200-1203

Scerif, G. & Karmiloff-Smith, A. (2005), The dawn of cognitive genetics? Crucial developmental caveats, *Trends in Cognitive Sciences* 9:126-135.

Schaffer, S. F. (2004), The X chromosome in population genetics, *Nature Reviews: Genetics* 5:43-51

Scharff, C. & Haesler, S. (2005), An evolutionary perspective on *FoxP2*: strictly for the birds? *Current Opinion in Neurobiology* 15:694-703.

Scharff, C. & White, S.A. (2004), Genetic components of vocal learning, *Annals of the New York Academy of Sciences* 1016:325-47.

Schmitt, R. (Ed.), 1989. *Compendium Linguarum Iranicarum*. Wiesbaden: Dr. Ludwig Reichert Verlag (especially articles by Josef Elfenbein on Balochi and Prods O. Skjaervø on Pashto)

Schwartz, J. H. (2004) Getting to know *Homo erectus*, *Science* 305:53-54

Schwartz, M., & Vissing, J. (2002), Paternal Inheritance of Mitochondrial DNA, *The New England Journal of Medicine*, 347:576-580

Seehausen, O. (2004), Hybridization and adaptive radiation, *TRENDS in Ecology and Evolution* 19:198-207.

Seeley, R.R., Stephens, T.D. & Tate, P. (2005), *Anatomy and physiology* (7th Edition), New York: McGraw Hill International Edition

Senut, B., Pickford, M., Gommery, D., Meind, P., Cheboie, K. & Coppens, Y. (2001), First hominid from the Miocene (Lukeino Formation, Kenya), *Comptes Rendus de l'Académie des Sciences - Series IIA - Earth and Planetary Science* 332:137-144

Serre, D., Manganey, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., Hofreiter, M., Possnert, G. & Pääbo, S. (2004), No evidence of Neandertal *mtDNA* contribution to eraly modern humans, *PloS Biology* 2:0313-0317

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46:561-576.

SHANNON, C. E. (1948), A Mathematical Theory of Communication, *The Bell System Technical Journal* 27:379–423, 623–656.

SHEEHAN, H. (1993), *Marxism and the Philosophy of Science: A Critical History*, Humanities Press International

SHEVOROSHKIN, V. (1999), *Nostratic languages: internal and external relationship*, In RENFREW, C. & NETTLE, D. (Eds.), pp.:75-92

SHU, W., CHO, J.Y., JIANG, Y., ZHANG, M., WEISZ, D., ELDER, G.A., SCHMEIDLER, J., DE GASPERI, R., SOSA, M.A., RABIDOU, D., SANTUCCI, A.C., PERL, D., MORRISEY, E., BUXBAUM, J.D. (2005), Altered ultrasonic vocalization in mice with a disruption in the *Foxp2* gene, *PNAS* 102:9643-9648.

SIMPSON, C. G. (1961), *Principles of animal taxonomy*, New York: Columbia University Press

SIMS-WILLIAMS, P. (1998), Genetics, linguistics, and prehistory: thinking big and thinking straight, *Antiquity* 72:505-527

SKELTON, P. (1993), *Evolution: a biological and palaeoantological approach*, The Open University

SMITH, K. (2003), *The Transmission of Language: models of biological and cultural evolution*, PhD Thesis, Theoretical and Applied Linguistics, University of Edinburgh

SMITH, K. (2004), The evolution of vocabulary, *Journal of Theoretical Biology* 228:127-142

SMITH, K. (2006), *The protolanguage debate: bridging the gap*?, In CANGELOSI, SMITH & SMITH (Eds.) (2006), pp. 315-322

SMITH, K., KIRBY, S., AND BRIGHTON, H. (2003) Iterated Learning: a framework for the emergence of language. *Artificial Life*, 9(4):371-386

SNA R PACKAGE: CARTER T. BUTTS (2006). *sna: Tools for Social Network Analysis*. R package version 1.1. http://erzuli.ss.uci.edu/R.stuff

SOKAL, R. R., ODEN, N. L. & THOMSON, B. A. (1992), Origins of the Indo-Europeans: Genetic evidence, *PNAS* 89:7669-7673

SPRINGER, M. S., STANHOPE, M. J., MADSEN, O. & DE JONG, W. W. (2004), Molecules consolidate the placental mammal tree, *Trends in Ecology and Evolution*, 19:430–438.

STAROSTIN, S. (1999), *Subgrouping of Nostratic: comments on Aharon Dolgopolsky's The Nostratic Macrofamily and Linguistic Palaeontology*, In RENFREW, C. & NETTLE, D. (Eds.), pp.:137-156

STAROSTIN, S. (2000), *Comparative-historical linguists and lexicostatistics*, In RENFREW, MCMAHON & TRASK (2000), pp 223-259

STEFANSSON, H., HELGASON, A., THORLEIFSSON, G., STEINTHORSDOTTIR, V., MASSON, G., BARNARD, J., BAKER, A., JONASDOTTIR, A., INGASON, A., GUDNADOTTIR, V. G., DESNICA, N., HICKS, A., GYLFASON, A., GUDBJARTSSON, D. F., JONSDOTTIR, G. M., SAINZ, J., AGNARSSON, K., BIRGISDOTTIR, B., GHOSH, S., OLAFSDOTTIR, A., CAZIER, J.-B., KRISTJANSSON, K., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R., KONG, A. & STEFANSSON, K. (2005) A common inversion under selection in Europeans, *Nature Genetics* 37:129-137

STEWART, J.R. (2005), The ecology and adaptation of Neanderthals during the non-analogue environment of Oxygen Isotope Stage 3, *Quaternary International* 137:35-46

STORM, P. (2001), The evolution of humans in Australasia from an environmental perspective, *Palaeogeography, Palaeoclimatology, Palaeoecology* 171:363-383

STRAIT, D. S. & GRINE, F. E. (2004), Inferring hominoid and early hominid phylogeny using craniodental characters: the role of fossil taxa, *Journal of Human Evolution* 47:399-452

STRINGER, C. (2002), Modern human origins: progress and prospects, *Philological Transactions of the Royal Society of London* B 357:563-579

STRINGER, C. & MCKIE, R. (1996), *African exodus*, London: Jonathan Cape

STRINGER, C. B. & ANDREWS, P. (1988), Genetic and fossil evidence for the origin of modern humans, *Science* 239:1263-1268

STROMSWOLD, K. (2001), The heritability of language: a review and metaanalysis of twin, adoption, and linkage studies, *Language* 77:647-723

SWADESH, M. (1952), Lexico-statistic dating of prehistoric ethnic contacts, *Proceedings of the American Philosophical Society* 96:452-463

SYKES, B. (2004), *The Seven Daughters of Eve*, Corgi Adult

TABACHNICK, B. G. & FIDELL, L. S. (2001), *Using multivariate statistics* (4th edition), Needham Heights, MA: Allyn & Bacon

TALLERMANN, M. (2006), *A holistic language cannot be stored, cannot be retrieved*, In CANGELOSI, SMITH & SMITH (Eds.) (2006), pp. 447-448.

TAMBETS, K., TOLK, H.-V., KIVISILD, T., METSPALU, E., PARIK, J., REIFLA, M., VOEVODA, M., DAMBA, L., BERMISHEVA, M., KHUSNUTDINOVA, E., GOLUBENKO, M., STEPANOV, V., PUZYREV, V., ESANGA, E., RUDAN, P., BECKMANN, L. & VILLEMS, R. (2002), *Complex signals for Population Expansions in Europe and Beyond*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:449-457

TATTERSALL, I. (2000), Palaeoanthropology: The last half-century, *Evolutionary Anthropology: Issues, News and Reviews* 9:2-16

TATTERSALL, I. & MOWBRAY, K. (2005) Species and paleoanthropology, *Theory in Biosciences* 123:371-379

TATTERSALL, I. & SCHWARTZ, J. H. (1999), Hominids and hybrids: The place of Neanderthals in human evolution, *Proceedings of the National Academy of Sciences of the USA* 96: 7117-7119

TEACHINGDEMOS R PACKAGE: GREG SNOW (2005). *TeachingDemos: Demonstrations for teaching and learning*. R package version 1.2.

TEMPLETON, A. R. (1993), The "Eve" hypothesis: A genetic critique and reanalysis, *American Anthropologist* 95:51-72

TEMPLETON, A. R. (1998), Human Races: A genetic and evolutionary perspective, *American Anthropologist* 100:632-650

TEMPLETON, A. R. (2002), Out of Africa again and again, *Nature* 416:45-51

TERAMITSU, I., KUDO, L. C., LONDON, S. E., GESCHWIND, D. H. & WHITE, S. A. (2004), Parallel *FoxP1* and *FoxP2* Expression in Songbird and Human Brain Predicts Functional Interaction, *The Journal of Neuroscience* 24:3152–31635

THE INTERNATIONAL HAPMAP CONSORTIUM (2005), A haplotype map of the human genome, *Nature* 437:1299-1320

THE SLI CONSORTIUM (SLIC), (2004) Highly Significant Linkage to the SLI1 Locus in an Expanded Sample of Individuals Affected by Specific Language Impairment, *American Journal of Human Genetics* 74:1225–1238

THE WORLD FACTBOOK (2005) https://www.cia.gov/cia/publications/factbook/index.html

THOMASON, S. (2000), *Linguistic areas and language history*. In GILBERS, D., NERBONNE, J. & SCHAEKEN, J. (Eds.), *Languages in Contact*. Rodopi: Amsterdam, pp. 311-327

THOMPSON, P. M., CANNON, T. D., NARR, K. L., VAN ERP, T., POUTANEN, V.-P., HUTTUNEN, M. *ET AL.* (2001), Genetic influences on brain structure, *Nature Neuroscience* 4:1253-1258

THORNE, A. G. & WOLPOFF, M. H. (2003), The multiregional evolution of humans, *Scientific American* 13:46-53

TOMIĆ, O.M. (2003), *The Balkan Sprachbund properties: An introduction to Topics in Balkan Syntax and Semantics*, LOT [Landelijke Onderzoekschool Taalwetenschap (Netherlands Graduate School of Linguistics)], Summer School 2003, Tilburg, The Netherlands, June 16-27, 2003 (online http://wwwlot.let.uu.nl/GraduateProgram/LotSchools/Summerschool2003/Tomic.pdf, August 2006)

TRANAH, G., CAMPTON, D. E. & MAY, B. (2004), Genetic evidence for hybridization of pallid and shovelnose sturgeon, *Journal of Heredity* 95:474-480.

TRASK, R. L. (1996), *Historical Linguistics*, London: Arnold

TRASK, R. L. (1999), *Why should a language have any relatives*?, In RENFREW, C. & NETTLE, D. (Eds.), pp.:157-178

TRAUTH, M. H., MASLIN, M. A., DEINO, A. & STRECKER, M. R. (2005) Late Cenozoic Moisture History of East Africa, *Science* 309:2051-2053

TRINKAUS, E. (1981), *Neanderthal limb proportions and cold adaptation*. In STRINGER, C. B., (ed.), *Aspects of Human Evolution*. London: Taylor & Francis, p. 187-219.

TRINKAUS, E. & ZILHÃO, J. (2003). *Phylogenetic implications*. In INSTITUTO PORTUGUES DE ARQUEOLOGIA (CORPORATE AUTHOR), J. ZILHÃO AND E. TRINKAUS (Eds.), *Portrait of the Artist As a Child: The Gravettian Human Skeleton From the Abrigo Do Lagar Velho and its Archaeological Context* (pp. 497-518). Oxbow Books Ltd.

TRINKAUS, E., MOLDOVAN, O., MILOTA, Ş., BÎLGĂR, A., SARCINA, L., ATHREYA, S., BAILEY, S. E., RODRIGO, R., MIRCEA, G., THOMAS, H., RAMSEY, C. B. & VAN DER PLICHT, J. (2003), An early modern human from the Peştera cu Oase, Romania, *PNAS* 100:11231-11236

TRIPACK R PACKAGE: FORTRAN CODE BY R. J. RENKA. R FUNCTIONS BY ALBRECHT GEBHARDT. WITH CONTRIBUTIONS FROM STEPHEN EGLEN, SERGEI ZUYEV AND DENIS WHITE (). *tripack: Triangulation of irregularly spaced data*. R package version 1.2-10.

TSAOUSIS, A. D., MARTIN, D. P., LADOUKAKIS, E. D. , POSADA, D. & ZOUROS, E. (2005), Widespread Recombination in Published Animal mtDNA Sequences, *Molecular Biology and Evolution* 22:925-933

TUCKER, A. N. & MPAAYEI, J. T. O. (1955). *A Maasai Grammar*. London: Longmans Green.

UNDERHILL, P. A. (2002), *Inference of Neolithic Populations Histories using Y-chromosome Haplotypes*, In BELLWOOD & RENFREW (Eds.) (2002), pp.:65-78

UPTON, G. & FINGLETON, B. (1985), *Spatial Data Analysis by Example*, Vol. I (Point pattern and quantitative data), John Wiley & Sons Ltd.

VALLADAS, H., JORON, J.L., VALLADAS, G., ARENSBURG, B., BAR-YOSEF, O., BELFER-COHEN, A., GOLDBERG, P., LAVILLE, H., MEIGNEN, L., RAK, Y., TCHERNOV, E., TILLIER, A.M. & VANDERMEERSCH, B. (1987), Thermoluminescence dates for the Neanderthal burial site at Kebara in Israel, *Nature* 330:159-160

VAN DEN BERGH, G. D, DE VOS, J. & SONDAAR, P. Y. (2001), The Late Quaternary palaeogeography of mammal evolution in the Indonesian archipelago, *Palaeogeography, Palaeoclimatology, Palaeoecology* 171:385-408

VAN DER LELY, H. K. J. (2005), Domain-specific cognitive systems: insight from Grammatical-SLI, *TRENDS in Cognitive Sciences* 9:53-59

VAN SCHAIK, C. P., DEANER, R. O. & MERRILL, M. Y. (1999), The conditions for tool use in primates: implications for the evolution of material culture, *Journal of Human Evolution* 36:719-741

VANHAEREN, W., D'ERRICO, F., STRINGER, C., JAMES, S. L., TODD, J. A. & MIENIS, H. K. (2006), Middle Plaeolithic Shell Beads in Israel and Algeria, *Science* 312:1785-1788.

VARGHA-KHADEM, F., GADIAN, D. G., COPP, A. & MISHKIN, M. (2005), *FOXP2* and the neuroanatomy of speech and language, *Nature Reviews* 6:131-138

VARGHA-KHADEM, F., WATKINS, K. E., PRICE, C. J., ASHBURNER, J., ALCOCK, K. J., CONNELLY, A., FRACKOWIAK, R. S. J., FRISTON, K. J., PEMBREY, M. E., MISHKIN, M. & PASSINGHAM, R. E. (1998), Neural basis of an inherited speech and language disorder, *PNAS* 95:12695-12700

VEGAN R PACKAGE: OKSANEN, J., KINDT, R., LEGENDRE, P. & O'HARA, R.B. (2006). *vegan: Community Ecology Package* version 1.8-2. http://cran.r-project.org/

VEKUA, A., LORDKIPANIDZE, D., RIGHTMIRE, G. P., AGUSTI, J., FERRING, R., MAISURADZE, G., MOUSKHELISHVILI, A., NIORADZE, M., DE LEON, M. P., TAPPEN, M., TVALCHRELIDZE, M., ZOLLIKOFER, C. (2002), A New Skull of Early Homo from Dmanisi, Georgia, *Science* 297:85-89

VERNESI, C., CARAMELLI, D., DUPANLOUP, I., BERTORELLE, G., LARI, M., CAPPELLINI, E., MOGGI-CECCHI, J., CHIARELLI, B., CASTRÌ, L.CASOLI, A., MALLEGNI, F., LALUEZA-FOX, C. & BARBUJANI, G. (2004), The Etruscans: A Population-Genetic Study, *American Journal of Human Genetics* 74:694-704

VIGILANT, L., STONEKING, M., HARPENDING, H., HAWKES, K. & WILSON A.C. (1991), African populations and the evolution of human mitochondrial DNA, *Science* 253:1503-1507

VILLMOARE, B. (2005) Metric and non-metric randomization methods, geographic variation, and the single-species hypothesis for Asian and African *Homo erectus*, *Journal Of Human Evolution* 49:680-701

VOIGHT, B. F., KUDARAVALLI, S., WEN, X. & PRITCHARD, J. K. (2006), A Map of Recent Positive Selection in the Human Genome, *PLoS Biology* 4:0446-0458

WAGNER, L. (2002), The heritability of language, *TRENDS in Congnitive Sciences* 6:198

WALKER, A. (2002), New Perspectives on the Hominids of the Turkana Basin, Kenya, *Evolutionary Anthropology, Suppl* 1:38 – 41

WALL, J. D. (2000). Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, 154, 1271-1279.

WALL, J. D. & PREZWORSKI, M. (2000), When did the human population size start

increasing?, *Genetics* 155:1865-1874

WALS Software: The World Atlas of Language Structures (Haspelmath, Dryer, Gil & Comrie (2005)) – *The Interactive Reference Tool*, developed by Bibiko, H.-J.

Walsh, B. (2004), *Multiple comparisons: Bonferroni Corrections and False Discovery Rates*, Lecture Notes (EEB 581, 14 May 2004), Department of Ecology and Evolutionary Biology, University of Arizona, online at http://nitro.biosci.arizona.edu/courses/EEB581-2004/handouts/Multiple.pdf (the 14th of August, 2006)

Wang, E.T., Kodama, G., Baldi, P. & Moyzis, R.K. (2006), Global landscape of recent inferred Darwinian selection for Homo sapiens, *PNAS* 103:135-140

Wang, H., Ambrose, S. H., Liu, C.-L. J. & Follmer, L. R. (1997), Paleosol Stable Isotope Evidence for Early Hominid Occupation of East Asian Temperate Environments, *Quaternary Research* 48:228-238

Watkins, K. E., Dronkers, N. F. & Vargha-Khadem, F. (2002), Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia, *Brain* 125:452-464.

Watkins, K. E., Vargha-Khadem, F., Ashburner, J., Passingham, R. E., Connelly, A., Friston, K. J., Frackowiak, R. S. J., Mishkin, M. & Gadian, D. G. (2002), MRI analysis of an inherited speech and language disorder: structural brain abnormalities, *Brain* 125:465-478.

Weaver, T. D. & Roseman, C. C. (2005), Ancient DNA, late Neandertal survival and modern-human – Neandertal genetic admixture, *Current Anthropology* 4:677-683

Webb, D.M. & Zhang, J. (2005), FoxP2 in song-learning birds and vocal-learning mammals, *Journal of Heredity* 96:212-216.

Weber, J., Czarnetzki, A. & Pusch, C. M. (2005) Comment on "The Brain of LB1, *Homo floresiensis*", *Science* 310:236 (Technical Comment)

Webster, R. & Olivier, M. A. (2001), *Geostatistics for Environmental Sciences*, Chichester, UK: John Wiley & Sons

Weidenreich, F. (1947a), Facts and speculations concerning the origin of Homo sapiens. *American Anthropologist* 49:187-203.

Weidenreich, F. (1947b), The trend of human evolution. *Evolution* 1:221-236.

Wescott, R. W. (1998), "What is Nostratic" (Table included in Bengtson, 1998), *Mother Tongue* 31:35

West-Eberhard, M. J. (2003), *Developmental plasticity and evolution*, Oxford, NY: Oxford University Press

White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G. & Howell, F. C. (2003), Pleistocene *Homo sapiens* from Middle Awash, Ethiopia, *Nature* 423:742-747

Wildavsky, A. (1997), *But Is It True?: Citizen's Guide to Environmental Health and Safety Issues*, Harvard University Press

Wilson, R. C. L., Drury, S. A. & Chapman, J. L. (2000), *The Great Ice Age: Climate Change and Life*, Routledge & The Open University: London

WOLPERT, L. (2001), *Principles of development* (2nd Edition), Oxford University Press

WOLPOFF, M. & CASPARI, R. (1997), *Race and human evolution: a fatal attraction*, Boulder, Colorado: Westview Press

WOLPOFF, M. H. & CASPARI, R. (2000), The many species of humanity, *Przegląd Antropologiczny/Anthroplogical Review* 63:3-17

WOLPOFF, M. H., HAWKS, J. & CASPARI, R. (2000), Multiregional, Not multiple origins, *American Journal of Physical Anthropology* 112:129-136

WOLPOFF, M. H., MANNHEIM, B., MANN, A., HAWKS, J., CASPARI, R., ROSENBERG, K. R., FRAYER, D. W., GILL, G. W. & CLARK, G. (2004), Why not the Neandertals?, *World Archaeology: Debates in World Archaeology* 36:527-546

WOLPOFF, M. H., WU, X. & THORNE, A. G. (1984), Modern *Homo sapiens* origins: a general theory of hominid evolution involving the fossil evidence from east Asia, In SMITH, F. H. & SPENCER F. (Eds.), *The origins of modern humans: a world survey of the fossil evidence*. NY: Alan R. Liss: 411-483

WOLPOFF, M.H., HAWKS, J., FRAYER, D. W. & HUNLEY, K. (2001), Modern Human Ancestry at the Peripheries: A Test of the Replacement Theory, *Science* 291:293-297

WRAY, A. (2000), *Holistic utterances in protolanguage: the link from primates to humans*. In KNIGHT, STUDDERT-KENNEDY & HURFORD (Eds.), pp. 285-302.

WRAY, A. (2002), *Dual processing in protolanguage: performance without competence*. In WRAY, A. (Ed.). *The Transition to Language*. Oxford: Oxford University Press. 113-137.

WRIGHT, S. (1931), Evolution in Mendelian populations, *Genetics* 16:97-159

WRIGHT, S. P. (1992). Adjusted P-values for simultaneous inference. *Biometrics*, 48:1005-1013.

WU, X. (2003), On the origin of modern humans in China, *Quaternary International* 117:131-140

XI, Z. (1996) *Vowel systems of the Manchu-Tungus languages of China*, PhD Thesis, Dept. Of Linguistics, University of Toronto (online http://r1.chass.utoronto.ca/twpl/pdfs/dissertations/Zhang.Xi.pdf)

YAMAUCHI, H. (2004), *Baldwinian Accounts of Language Evolution*, PhD Thesis, Theoretical and Applied Linguistics, The University of Edinburgh. Online http://www.ling.ed.ac.uk/~hoplite/publications/yamauchi-PhD.pdf (November 2006).

YORK, R. (2005), *Homo Floresiensis* and Human Equality: Enduring Lessons from Stephen Jay Gould, *Monthly Review* 56 (10), March 2005.

YOUNG, W. P., OSTBERG, C. O., KEIM, P. & THORGAARD, G. H. (2001), Genetic characterization of hybridization and introgression between anadromous rainbow trout (*Onchorhynchus mykiss irideus*) and coastal cutthroat trout (*O. clarki clarki*), *Molecular Ecology* 10:921-931.

YU, N., FU, Y.-X. & LI, W.-H. (2002), DNA Polymorphism in a Worldwide Sample of Human X Chromosomes, *Molecular Biology and Evolution* 19:2131-2141

YU, N., JENSEN-SEAMAN, M. I., CHEMNICK, L., KIDD, J. R., DEINARD, A. S., RYDER, O., KIDD, K. K. & LI, W.-H. (2003), Low nucleotide diversity in chimpanzees and bonobos, *Genetics* 164:1511-1518

ZHANG, J., WEBB, D.M. & PODLAHA, O. (2002), Accelerated protein evolution and origins of human-specific features: *Foxp2* as an example, Genetics 162:1825-1835.

ZHU, R. X., HOFFMAN, K. A., POTTS, R., DENG, C. L., PAN, Y. X., GUO, B., SHI, C. D., GUO, Z. T., YUAN, B. Y., HOU, Y.M. & HUANG, W. W. (2001), Earliest presence of humans in northeast Asia, *Nature* 413:413-417

ZIĘTKIEWICZ, E., YOTOVA, V., GEHL, D., WAMBACH, T., ARRIETA, I., BATZER, M., COLE, D. E. C., HECHTMAN, P., KAPLAN, F., MODIANO, D., MOISAN, J.-P., MICHALSKI, R. & LABUDA, D. (2003), Haplotypes in the *Dystrophin* DNA segment point to a mosaic origin of modern human diversity, *American Journal of Human Genetics* 73:994-1015

ZILHÃO, J. & TRINKAUS, E. (2003a). *Historical implications*. In INSTITUTO PORTUGUES DE ARQUEOLOGIA (CORPORATE AUTHOR), J. ZILHÃO AND E. TRINKAUS (Eds.), *Portrait of the Artist As a Child: The Gravettian Human Skeleton From the Abrigo Do Lagar Velho and its Archaeological Context* (pp. 542-558). Oxbow Books Ltd.

ZILHÃO, J. & TRINKAUS, E. (2003b). *Social implications*. In INSTITUTO PORTUGUES DE ARQUEOLOGIA (CORPORATE AUTHOR), J. ZILHÃO AND E. TRINKAUS (Eds.), *Portrait of the Artist As a Child: The Gravettian Human Skeleton From the Abrigo Do Lagar Velho and its Archaeological Context* (pp. 519-541). Oxbow Books Ltd.

ZIV, J. & LEMPEL, A. (1977), A universal algorithm for sequential data compression, *IEEE Transactions on Information Theory* 23:337-342

ZUCKERKANDL, E. & PAULING, L. (1962), *Molecular disease, evolution, and genetic heterogeneity*, In KASHA, M. & PULLMAN, B. (Eds.), *Horizons in Biochemistry*, Academic Press: New York, pp. 189–225.

ZUCKERKANDL, E. & PAULING, L. (1965), *Evolutionary divergence and convergence in proteins*, In BRYSON, V. & VOGEL, H. J. (Eds.), *Evolving Genes and Proteins*, Academic Press: New York, pp. 97–166.

ZVELEBIL, M. (2002), *Demography and dispersal of first farming populations at the Mesolithic-Neolithic transition: linguistic implications*, In Bellwood & Renfrew (Eds.) (2002), pp.:379–394.